To appear in *Optimization Methods & Software*
Vol. 00, No. 00, Month 20XX, 1–35

RESEARCH ARTICLE

Complexity of first order inexact Lagrangian and penalty methods for conic convex programming

I. Necoara[a*] and A. Patrascu[a] and F. Glineur[b]

[a] *Automatic Control and Systems Engineering Department, University Politehnica Bucharest, 060042 Bucharest, Romania.* [b] *Center for Operations Research and Econometrics, Catholic University of Louvain, B-1348 Louvain-la-Neuve, Belgium*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

In this paper we present a complete iteration complexity analysis of inexact first order Lagrangian and penalty methods for solving cone constrained convex problems that have or may not have optimal Lagrange multipliers that close the duality gap. We first assume the existence of optimal Lagrange multipliers and study primal-dual first order methods based on inexact information and augmented Lagrangian smoothing or Nesterov type smoothing. For inexact (fast) gradient augmented Lagrangian methods we derive a total computational complexity of $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ projections onto a simple primal set in order to attain an $\epsilon-$optimal solution of the conic convex problem. For the inexact fast gradient method combined with Nesterov type smoothing we derive computational complexity $\mathcal{O}\left(\frac{1}{\epsilon^{3/2}}\right)$ projections onto the same set. Then, we assume that optimal Lagrange multipliers for the cone constrained convex problem might not exist, and analyze the fast gradient method for solving penalty reformulations of the problem. For the fast gradient method combined with penalty framework we also derive a total computational complexity of $\mathcal{O}\left(\frac{1}{\epsilon^{3/2}}\right)$ projections onto a simple primal set to attain an $\epsilon-$optimal solution for the original problem.

Keywords: conic convex problems, smooth (augmented) dual functions, penalty functions, (augmented) dual first order methods, penalty fast gradient methods, approximate primal solution, computational complexity.

AMS Subject Classification: 90C25; 90C46; 68Q25; 65K05.

1. Introduction

Many recent engineering and economical applications can be posed as large-scale conic convex problems and thus the interest for scalable algorithms with inexpensive iterations is continuously increasing. For instance, in the recent optimization literature, first order methods gained much attention since they present cheap iterations and are usually adequate for large-scale convex setting. In the constrained case, when there are conic complicated constraints, many first order algorithms are combined with duality or penalty strategies. For example, in [8] various smooth and nonsmooth formulations are provided

*Corresponding author, email: ion.necoara@acse.pub.ro. The research leading to these results has received funding from UEFISCDI Romania, PNII-RU- TE, project MoCOBiDS, no. 176/01.10.2015. It also presents research results of the Belgian Network DYSCO funded by the Interuniversity Attraction Poles Programme initiated by the Belgian State, and of the Concerted Research Action programme supported by the Federation Wallonia-Brussels, no. ARC 14/19-060. Support from two WBI-Romanian Academy grants is also acknowledged. The authors thank Prof. Yu. Nesterov for inspiring discussions.

for cone programming, and through application of first order methods (e.g. fast gradient or mirror descent) on the corresponding reformulations of the optimality conditions as optimization problems, an ϵ-optimal solution is obtained in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ projections onto a simple primal set. For conic constrained convex problems, quadratic penalty strategies are combined with fast gradient method in [9]. Under the assumptions of smooth objective function and existence of a finite optimal Lagrange multiplier, the first order quadratic penalty method in [9] requires $\mathcal{O}(\frac{1}{\epsilon^2})$ fast gradient iterations. Moreover, using a regularization of the original problem with a strongly convex term, this method requires $\mathcal{O}(\frac{1}{\epsilon}\log\left(\frac{1}{\epsilon}\right))$ fast gradient iterations. Recently, other first order augmented Lagrangian methods are presented in [1, 10, 16] and computational complexity estimates of order $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ are obtained for smooth problems with bounded optimal Lagrange multipliers. First order methods are also combined with duality and Nesterov type smoothing in [3–5, 7, 15, 23, 24, 27] and convergence rates of order $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ in terms of dual gradient evaluations are derived. Another interesting approach relies on reformulation of conic constrained programming problems into a monotone variational inequality and then designing various algorithms for solving these inequalities. This approach can be found in [18, 20], where different primal-dual methods are devised for solving the variational inequality under the boundedness assumption of the primal and dual feasible sets. Recently, the boundedness condition has been eliminated in [11].

Motivation. However, the following issues can be identified in the existing literature:
(a) Most of the existing papers on dual first order methods combined with smoothing techniques derive rate of convergence results in terms of outer iterations (number of dual gradient evaluations). However, we will show (see e.g., Theorem 3.4) that one might choose an appropriate value of the smoothing parameter such that after a single outer iteration an $\epsilon-$solution can be obtained. Thus, convergence rates in terms of outer iterations are not relevant in this case and it is natural to analyze the overall complexity of these methods that also take into account the inner iterations (e.g., number of projections onto the primal feasible set or number of matrix-vector multiplications).
(b) Moreover, from our knowledge, there is no complete analysis in the optimization literature regarding the overall complexity of inexact dual first order methods based on augmented Lagrangian smoothing and Nesterov smoothing and clarifying which smoothing approach has a better behavior.
(c) Finally, all the papers on Lagrangian and penalty methods mentioned above make the strong assumption that there exists an optimal Lagrange multiplier for the primal convex problem that closes the duality gap. This property is usually guaranteed through a Slater type condition, which in the large-scale setting is very difficult to check computationally or even might not hold. Recently, Nesterov developed in [22] subgradient methods for nonsmooth convex problems with functional constraints without this assumption on the existence of an optimal Lagrange multiplier and proved that an ϵ-optimal point can be attained after $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ subgradient evaluations for either the objective function or for a functional constraint. Nesterov also asks in [22] whether it is possible to improve this convergence rate result under additional smooth assumptions on the objective function and functional constraints.

Contributions. These issues motivate our work here. In this paper we present a complete iteration complexity analysis of inexact first order Lagrangian and penalty methods for solving cone constrained convex problems that have or may not have optimal Lagrange multipliers that close the duality gap. In the first part of our paper we assume the existence of optimal Lagrange multipliers and we derive overall complexity of primal-dual first order methods based on the inexact oracle framework [6] and augmented Lagrangian

smoothing [25] or Nesterov type smoothing [5, 19]. Although we obtain in some cases similar complexity results with those found in the literature, our analysis based on the inexact oracle framework is simpler, intuitive and more elegant, opening various possibilities for extensions to more complex optimization models. Moreover, in some optimality criteria our computational complexities are significantly better than those found in the existing literature. These better complexities are achieved through the new first order inexact oracles for augmented Lagrangian (Nesterov) smoothing derived in Theorem 3.2 (Theorem 3.11) that improve substantially those in [6]. In the second part we assume that the conic constrained convex problem might not admit an optimal Lagrange multiplier. In this case, we combine the fast gradient method with penalty strategies and derive computational complexity certifications for such methods which consistently improves those given in [22] for the nonsmooth case. Thus, our results cover the particular case when the Slater condition does not hold or it is difficult to check for large-scale conic convex problems and answer positively to Nesterov's question. To the best of our knowledge, this paper present one of the first computational complexity results for first order penalty methods for convex problems when optimal Lagrange multipliers do not exist. More explicitly, our contributions are:

(i) First, we assume that we have optimal Lagrange multipliers that close the duality gap for the cone constrained convex problem with simple or smooth objective function. We provide new computational complexity results on the dual first order augmented Lagrangian methods, where the main complexity bounds show that, in order to obtain an $\epsilon-$optimal solution for the original problem, the inexact (fast) gradient augmented Lagrangian algorithms have to perform $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ total projections onto the simple primal feasible set and feasible cone.
(ii) We combine in a novel fashion Nesterov smoothing technique and inexact fast gradient method for solving cone constrained optimization problems with possibly unbounded feasible cone. We show that, in order to obtain an ϵ-optimal solution, fast gradient method with inexact information performs $\mathcal{O}\left(\frac{1}{\epsilon^{3/2}}\log\left(\frac{1}{\epsilon}\right)\right)$ projections onto the simple primal feasible set and $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ projections onto the feasible cone. Thus, our work shows that inexact fast gradient method based on Nesterov smoothing has worse overall complexity than the one based on augmented Lagrangian smoothing.
(iii) Then, we eliminate the assumption that there exists some optimal Lagrange multiplier for the cone constrained convex problem and we analyze the computational complexity of fast gradient penalty methods. If the objective function is smooth, then we prove that in order to obtain an ϵ-optimal solution for the original problem we need to perform $\mathcal{O}\left(\frac{1}{\epsilon^{3/2}}\right)$ total projections onto the simple primal feasible set. Through an example, we also show that our bounds are tight.

Notations. We denote $\bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$. For $u, v \in \mathbb{R}^n$, we consider scalar product $\langle u, v\rangle = u^T v$ and Euclidean norm $\|u\| = \sqrt{u^T u}$. Further, $[u]_U$ denotes the projection of u onto nonempty closed convex set U and $\mathrm{dist}_U(u) = \|u - [u]_U\|$ its distance to U. Moreover, we use notation $\mathcal{N}_U(u)$ for the normal cone of the convex set U at $u \in U$ defined by $\mathcal{N}_U(u) = \{t \in \mathbb{R}^n : \langle t, u - v\rangle \geq 0 \quad \forall v \in U\}$. We also use notation $\mathcal{B}_r(x) = \{z \in \mathbb{R}^n | \ \|z - x\| \leq r\}$. For a matrix $G \in \mathbb{R}^{m\times n}$ we use $\|G\|$ for the spectral norm.

2. Problem formulation

In this paper we consider the following cone constrained convex optimization problem:

$$f^* = \min_{u \in U} \quad f(u) \qquad \text{s.t.} \quad Gu + g \in \mathcal{K}, \tag{1}$$

where $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a proper, closed, convex function, $U \subseteq \mathrm{dom} f$ is a nonempty closed, convex set, $G \in \mathbb{R}^{m \times n}$ and $\mathcal{K} \subseteq \mathbb{R}^m$ is a nonempty, closed, convex cone, having its polar cone $\mathcal{K}^* = \{v \in \mathbb{R}^m : \langle v, \kappa \rangle \leq 0 \quad \forall \kappa \in \mathcal{K}\}$. We denote $U^* \subseteq \mathbb{R}^n$ the optimal set of the above problem. Note that our formulation and results can be extended to general normed vector spaces. The following assumptions are valid throughout the paper:

Assumption 2.1 *Objective function f is strongly convex with constant $\sigma_f \geq 0$:*

$$f\left(\alpha u + (1-\alpha) v\right) \leq \alpha f(u) + (1-\alpha) f(v) - \frac{\sigma_f \alpha (1-\alpha)}{2} \|u - v\|^2 \;\; \forall u, v \in dom f, \;\; \alpha \in [0,\, 1].$$

Note that Assumption 2.1 with $\sigma_f = 0$ is equivalent with convexity of function f.

Assumption 2.2 *(i) The feasible set U and the cones $\mathcal{K}$ and $\mathcal{K}^*$ are closed, convex and simple (e.g., the projection onto these sets can be obtained in closed form).*
(ii) The convex set U is bounded, i.e. exists $D_U < \infty$ such that $\max_{u,v \in U} \|u - v\| \leq D_U$.

Note that these assumptions are standard in the context of first order Lagrangian and penalty methods for conic convex programming, see e.g. [1, 9, 10, 18, 19, 24]. Further, convex function $h : \mathbb{R}^n \to \bar{\mathbb{R}}$, with $U \subseteq \mathrm{dom} h$, is called *simple* if the optimal solution of the following problem can be efficiently obtained (e.g., in closed form):

$$\min_{u \in U} \; h(u) + \frac{1}{2\mu} \|u - z\|^2 \qquad \forall \mu > 0 \;\text{ and }\; z \in \mathbb{R}^n.$$

In this paper we assume that the convex objective function f is either simple or has Lipschitz continuous gradient with constant $L_f > 0$ and $\mathrm{dom} f = \mathbb{R}^n$, i.e.:

$$0 \leq f(y) - \left(f(x) + \langle \nabla f(x), y - x \rangle\right) \leq \frac{L_f}{2} \|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n.$$

Our goal is to find an approximate solution for the optimization problem (1). Thus, we introduce the following definition used in the rest of the paper:

Definition 1 Given the desired accuracy $\epsilon > 0$, the primal point $u_\epsilon \in U$ is an *ϵ-optimal solution* for the cone constrained convex problem (1) if it satisfies:

$$|f(u_\epsilon) - f^*| \leq \epsilon \quad \text{and} \quad \mathrm{dist}_{\mathcal{K}}(Gu_\epsilon + g) \leq \epsilon.$$

2.1 *A framework for inexact first order methods*

Since the main algorithm in this paper is the Nesterov fast gradient method [21], we further introduce an inexact algorithmic framework based on the method in [2, 26], which will be subsequently called in various ways. Therefore, consider the following general convex constrained optimization problem with composite objective function:

$$F^* = \min_{z \in Q} \; F(z) \qquad \left(= \phi(z) + \psi(z)\right), \tag{2}$$

where $Q \subseteq \mathbb{R}^n$ is a simple, convex set, $\phi : \mathbb{R}^n \to \mathbb{R}$ is a convex function with Lipschitz continuous gradient of constant $L_\phi > 0$ and $\psi : \mathbb{R}^n \to \bar{\mathbb{R}}$ is a simple, closed, convex function. Using the definition from [6], given $\delta \geq 0$ and $L > 0$, we assume that the smooth function ϕ is equipped with a *first order inexact* (δ, L)*-oracle*, i.e. for any $y \in Q$ we can compute an approximate function value $\phi_{\delta,L}(y)$ and an approximate gradient $\nabla\phi_{\delta,L}(y)$ such that the following inequalities hold:

$$0 \leq \phi(x) - (\phi_{\delta,L}(y) + \langle \nabla\phi_{\delta,L}(y), x - y\rangle) \leq \frac{L}{2}\|x - y\|^2 + \delta \qquad \forall x \in Q. \tag{3}$$

Next, we introduce the Inexact Composite Fast Gradient (ICFG) method for solving the composite optimization problem (2) using approximate function values and gradients $(\phi_{\delta,L}(y), \nabla\phi_{\delta,L}(y))$ satisfying the first order (δ, L)-oracle given in (3):

Algorithm ICFG (ϕ, ψ, δ, L)
Give $z^0 = w^1 \in \mathbb{R}^n$ and $\theta_1 = 1$. For $k \geq 1$ do:

(1) Compute the pair $(\phi_{\delta,L}(w^k), \nabla\phi_{\delta,L}(w^k))$ satisfying (3). Update:
(2) $z^k = \arg\min\limits_{z\in Q} \langle \nabla\phi_{\delta,L}(w^k), z - w^k\rangle + \frac{L}{2}\|z - w^k\|^2 + \psi(z)$
(3) $w^{k+1} = z^k + \frac{\theta_k - 1}{\theta_{k+1}}(z^k - z^{k-1})$
(4) If a stopping criterion holds, then STOP and **return**: (z^k, w^k).

Note that if we update $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$ and additionally we consider $\delta = 0$ and $L = L_\phi$, then we recover the well-known FISTA scheme which has been analyzed for the first time in [2] and subsequently extended in different variants in [21, 26]. On the other hand, if we take $\theta_k = 1$ for all $k \geq 1$ and $\delta = 0$, then $z^k = w^{k+1}$ and we recover the ISTA scheme also developed in [2] and extended in [12, 17, 21]. Using the same reasoning as in [6], we provide in the next theorem the rate of convergence of Algorithm **ICFG** for composite optimization problem (2) endowed with a first order inexact (δ, L)-oracle (3). First, let us denote by z^* an optimal solution of the composite convex problem (2).

THEOREM 2.3 *[2, 6] Let sequences $(z^k, w^k)_{k\geq 0}$ be generated by Algorithm **ICFG** (ϕ, ψ, δ, L) for solving composite problem (2) endowed with a first order inexact (δ, L)-oracle. Then, we have the following convergence rates in terms of function values:*
(i) If we define the average sequence $\hat{z}^k = \frac{1}{k}\sum\limits_{i=0}^{k-1} z^{i+1}$ and $\theta_k = 1$ for all $k \geq 1$, then $\hat{z}^k$ has the following sublinear convergence rate in terms of function values:

$$F(\hat{z}^k) - F^* \leq \frac{L\|z^0 - z^*\|^2}{2k} + \delta.$$

(ii) If we update $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$, then the last iterate z^k has the following sublinear convergence rate in terms of function values:

$$F(z^k) - F^* \leq \frac{2L\|z^0 - z^*\|^2}{(k+1)^2} + k\delta.$$

3. Inexact first order Lagrangian methods

In this section we analyze the computational complexity of inexact first order Lagrangian methods for solving the cone constrained convex problem (1). Since we use the dual framework, we require the following standard assumption for dual algorithms, valid only in this Section 3:

ASSUMPTION 3.1 *There exists a Lagrange multiplier $x^* \in \mathcal{K}^*$ for the conic convex problem* (1) *that closes the duality gap.*

Assumption 3.1 implies the existence of a bounded optimal Lagrange multiplier, that is $\|x^*\| < \infty$, and it holds for (1) whenever a Slater type condition is valid, i.e. there exists $\bar{u} \in \mathrm{relint}(U)$ such that $G\bar{u} + g \in \mathrm{relint}(\mathcal{K})$.

3.1 *Preliminaries*

The strongly convex case, i.e. when the objective function f in problem (1) satisfies Assumption 2.1 with $\sigma_f > 0$, has been extensively studied in the literature, see e.g. [12, 13, 24, 27]. Thus, in the rest of our paper, unless it is explicitly stated, we assume that the function f is convex, i.e. it satisfies Assumption 2.1 with $\sigma_f = 0$. In the general convex case, the dual function, denoted d, is nonsmooth, and thus dual first order methods, such as Algorithm **ICFG**, cannot be applied. In order to be able to apply dual first order algorithms, our approach relies on the combination between smoothing techniques and duality. First, we introduce some notations. We note that the problem (1) can be reformulated equivalently as:

$$\min_{u,s} \; f(u) \quad \text{s.t.} \;\; u \in U, \;\; s \in \mathcal{K}, \;\; Gu + g = s. \tag{4}$$

Thus, the Lagrangian and the dual function of the convex problem (4) are given by:

$$\mathcal{L}(u,s,x) = f(u) + \langle x, Gu + g - s \rangle \quad \text{and} \quad d(x) = \min_{u \in U, s \in \mathcal{K}} \mathcal{L}(u,s,x).$$

Assumption 3.1 states that there exists a Lagrange multiplier $x^* \in \mathcal{K}^*$ such that $f^* = d(x^*)$ and thus the convex problem (1) is equivalent with solving the dual formulation:

$$f^* = \max_{x \in \mathbb{R}^m} d(x). \tag{5}$$

We denote with X^* the set of optimal solutions of the dual problem (5). Various dual subgradient schemes have been developed for solving (5) with ϵ accuracy, with overall complexities of order $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ [17, 22]. However, under additional mild assumptions, we aim in this paper at improving the iteration complexity required for solving the conic optimization problem (1) using the dual formulation. First, we rewrite the dual function d in a novel way as a composite function:

$$d(x) = \underbrace{\min_{u \in U} \left[f(u) + \langle x, Gu + g \rangle \right]}_{d_U(x)} + \underbrace{\min_{s \in \mathcal{K}} \langle -s, x \rangle}_{d_{\mathcal{K}}(x)} = d_U(x) + d_{\mathcal{K}}(x). \tag{6}$$

The function $d_{\mathcal{K}}(x)$ is the support function of the cone $\mathcal{K}$ and, by the definition of the polar cone $\mathcal{K}^*$, also represents the indicator function of $\mathcal{K}^*$. From our knowledge there are two widely known smoothing strategies to obtain an approximate dual function with Lipschitz continuous gradient. They are based on the following modified Lagrangian and dual functions:
(i) Augmented Lagrangian smoothing [1, 10, 16, 24]:

$$\mathcal{L}_\mu^{\mathrm{ag}}(u,s,x) = f(u) + \langle x, Gu + g - s\rangle + \frac{\mu}{2}\|Gu + g - s\|^2$$
$$d_\mu^{\mathrm{ag}}(x) = \min_{u \in U, s \in \mathcal{K}} \mathcal{L}_\mu^{\mathrm{ag}}(u,s,x).$$

(ii) Nesterov smoothing [3, 4, 15, 19, 23, 27]:

$$\mathcal{L}_\mu^{\mathrm{ns}}(u,s,x) = f(u) + \langle x, Gu + g - s\rangle + \frac{\mu}{2}(\|u\|^2 + \|s\|^2)$$
$$d_\mu^{\mathrm{ns}}(x) = \min_{u \in U, s \in \mathcal{K}} \mathcal{L}_\mu^{\mathrm{ns}}(u,s,x).$$

Note that, following the reasoning from [15, 19, 23], the Nesterov smoothing approximation $d_\mu^{\mathrm{ns}}(x)$ requires the boundedness of the primal feasible set $\mathcal{K} \times U$. Thus, the general convex cone $\mathcal{K}$ induces difficulties in using this strategy. In Section 3.4 we present a modified Nesterov smoothing technique which is able to cope with linear conic constraints and unbounded feasible cone $\mathcal{K}$ based on the new composite reformulation (6).

3.2 *Inexact first order methods for augmented Lagrangian smoothing*

In this section, we analyze the iteration complexity of the inexact first order methods for augmented Lagrangian smoothing, under Assumption 2.1 with $\sigma_f = 0$, Assumption 2.2 and Assumption 3.1. The inexact gradient Lagrangian method is equivalent with the classical augmented Lagrangian algorithm, namely the application of the inexact gradient method on the augmented dual function. The second first order Lagrangian method we analyze is the inexact fast gradient Lagrangian method, which is based on the application of the fast gradient method on the augmented dual function. We start with the classical augmented Lagrangian setting, i.e. we define [25]:

$$\mathcal{L}_\mu^{\mathrm{ag}}(u,x) = f(u) + \frac{\mu}{2}\mathrm{dist}_{\mathcal{K}}\left(Gu + g + \frac{1}{\mu}x\right)^2 - \frac{1}{2\mu}\|x\|^2 \;\text{ and }\; d_\mu^{\mathrm{ag}}(x) = \min_{u \in U} \mathcal{L}_\mu^{\mathrm{ag}}(u,x).$$

Note that, the augmented dual function represents a pure Moreau approximation of the original dual function:

$$d_\mu^{\mathrm{ag}}(x) = \max_{z \in \mathbb{R}^m} d(z) - \frac{1}{2\mu}\|z - x\|^2 = \max_{z \in \mathcal{K}^*} d_U(z) - \frac{1}{2\mu}\|z - x\|^2.$$

Further, we observe that partial gradient of $\mathcal{L}_\mu^{\mathrm{ag}}$ w.r.t. x is given by:

$$\nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u,x) = Gu + g - \left[Gu + g + \frac{1}{\mu}x\right]_{\mathcal{K}}.$$

For any $x \in \mathbb{R}^m$ we denote a primal exact solution by $u^*_\mu(x) \in \arg\min_{u \in U} \mathcal{L}^{\text{ag}}_\mu(u,x)$. It is well-known, see e.g. [25], that the gradient of augmented dual function $d^{\text{ag}}_\mu(x)$ satisfies:

$$\nabla d^{\text{ag}}_\mu(x) = Gu^*_\mu(x) + g - \left[Gu^*_\mu(x) + g + \frac{1}{\mu} x \right]_{\mathcal{K}},$$

and additionally it is Lipschitz continuous with constant $L_\text{d} = \frac{1}{\mu}$. Moreover, the resulting augmented dual problem, given by:

$$f^* = \max_{x \in \mathbb{R}^m} \; d^{\text{ag}}_\mu(x), \quad \text{satisfies} \quad X^* = \arg\max_{x \in \mathbb{R}^m} \; d^{\text{ag}}_\mu(x).$$

Usually, it is difficult to compute in most of the practical applications the optimal solution $u^*_\mu(x)$ of the inner problem $\min_{u \in U} \mathcal{L}^{\text{ag}}_\mu(u,x)$ and we can obtain only an approximate solution. Assume that we solve inexactly the inner problem and obtain an approximate solution $u_\mu(x) \in U$ which, for a given accuracy $\delta > 0$, satisfies:

$$0 \leq \mathcal{L}^{\text{ag}}_\mu(u_\mu(x), x) - d^{\text{ag}}_\mu(x) \leq \delta \quad \forall x \in \mathbb{R}^m. \tag{7}$$

Then, we can construct a first order inexact oracle for the augmented dual function:

THEOREM 3.2 *Let $\mu, \delta > 0$, then we have the following first order inexact $(3\delta, 2L_d)$-oracle for the augmented dual function d^{ag}_μ:*

$$0 \leq \mathcal{L}^{ag}_\mu(u_\mu(y), y) + \langle \nabla_x \mathcal{L}^{ag}_\mu(u_\mu(y), y), x - y \rangle - d^{ag}_\mu(x) \leq \frac{2L_d}{2} \|x - y\|^2 + 3\delta, \tag{8}$$

for all $x, y \in \mathbb{R}^m$, where the approximate solution $u_\mu(y)$ satisfies (7) and $L_d = \frac{1}{\mu}$.

Proof. For the left hand side inequality of (8), we observe that:

$$\mathcal{L}^{\text{ag}}_\mu(u_\mu(y), y) + \langle \nabla_x \mathcal{L}^{\text{ag}}_\mu(u_\mu(y), y), x - y \rangle \geq \mathcal{L}^{\text{ag}}_\mu(u_\mu(y), x) \geq \min_{u \in U} \; \mathcal{L}^{\text{ag}}_\mu(u, x) = d^{\text{ag}}_\mu(x).$$

For the right hand side inequality of (8), note that for any fixed $u \in U$ the function $h(x) = \mathcal{L}^{\text{ag}}_\mu(u,x) - d^{\text{ag}}_\mu(x)$ has Lipschitz gradient with constant $L_\text{h} = 2/\mu$ and $h(x) \geq 0$ for all $x \in \mathbb{R}^m$. Therefore, using the notation $L_\text{d} = 1/\mu$, we have:

$$h(x) - \min_{x \in \mathbb{R}^m} h(x) \geq \frac{1}{2L_\text{h}} \|\nabla h(x)\|^2 = \frac{1}{4L_\text{d}} \|\nabla_x \mathcal{L}^{\text{ag}}_\mu(u,x) - \nabla d^{\text{ag}}_\mu(x)\|^2 \qquad \forall u \in U.$$

Taking $u = u_\mu(x)$ and using the definition of $u_\mu(x)$, we have $h(x) - \min_{x \in \mathbb{R}^m} h(x) \leq h(x) \leq \delta$ and obtain the following approximate gradient relation:

$$\|\nabla_x \mathcal{L}^{\text{ag}}_\mu(u_\mu(x), x) - \nabla d^{\text{ag}}_\mu(x)\| = \|Gu_\mu(x) - Gu^*_\mu(x)\| \leq \sqrt{4\delta L_\text{d}}. \tag{9}$$

From the Lipschitz continuity of ∇d^{ag}_μ, (7) and (9), we have that for any $x, y \in \mathbb{R}^m$ the

following relations hold:

$$
\begin{aligned}
d_\mu^{\text{ag}}(x) &\geq d_\mu^{\text{ag}}(y) + \langle \nabla d_\mu^{\text{ag}}(y), x-y\rangle - \frac{L_{\text{d}}}{2}\|x-y\|^2 \\
&\overset{(7)}{\geq} \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y) + \langle \nabla d_\mu^{\text{ag}}(y), x-y\rangle - \frac{L_{\text{d}}}{2}\|x-y\|^2 - \delta \\
&= \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y) + \langle \nabla_x \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y), x-y\rangle - \frac{L_{\text{d}}}{2}\|x-y\|^2 - \delta \\
&\qquad\qquad + \langle \nabla d_\mu^{\text{ag}}(y) - \nabla_x \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y), x-y\rangle \\
&\geq \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y) + \langle \nabla_x \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y), x-y\rangle - \frac{L_{\text{d}}}{2}\|x-y\|^2 - \delta \\
&\qquad\qquad - \|\nabla d_\mu^{\text{ag}}(y) - \nabla_x \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y)\|\|x-y\| \\
&\overset{(9)}{\geq} \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y) + \langle \nabla_x \mathcal{L}_\mu^{\text{ag}}(u_\mu(y),y), x-y\rangle - \frac{L_{\text{d}}}{2}\|x-y\|^2 - \delta - \sqrt{4\delta L_{\text{d}}}\|x-y\|.
\end{aligned}
$$

On the other hand, for any positive pair of constants (t,α) we have: $\alpha t \leq \frac{t^2}{2} + \frac{\alpha^2}{2}$. Thus, taking $t = \sqrt{L_{\text{d}}}\|x-y\|$ and $\alpha = 2\sqrt{\delta}$ in the previous inequalities, we obtain the right hand side inequality of the theorem. ■

The relation (8) implies that the augmented dual function d_μ^{ag} is smooth and is equipped with a first order inexact $(3\delta, 2L_{\text{d}})$-oracle having $\phi_{\delta,L}(x) = \mathcal{L}_\mu^{\text{ag}}(u_\mu(x),x)$ and $\nabla\phi_{\delta,L}(x) = \nabla_x \mathcal{L}_\mu^{\text{ag}}(u_\mu(x),x) = Gu_\mu(x) + g$. It is important to note that many previous results on augmented Langragian methods require solving the inner problem with much higher inner accuracy of order $\mathcal{O}(\delta^2)$ (see e.g. [1, 6, 10, 16, 24]), i.e.:

$$
\mathcal{L}_\mu^{\text{ag}}(u_\mu(x),x) - d_\mu^{\text{ag}}(x) \leq \mathcal{O}(\delta^2).
$$

It is obvious that our approach here is less conservative, imposing to solve the inner problem with less inner accuracy of order δ as in (7). As we will see in the sequel, this new and important result from Theorem 3.2 will have a huge impact on the computational complexity of our methods compared to those given in the previous papers. In particular, the first order inexact oracle derived in [6] for augmented Lagrangian dual function is more conservative than the one from Theorem 3.2 and thus, its direct application will lead to much worse computational complexities than the ones we obtained in the present paper based on Theorem 3.2.

Given the pair $(x^k, y^k)_{k\geq 0}$ generated by Algorithm **ICFG**, in the following two sections we provide complexity estimates related to the convergence of the average primal point $(\hat{u}^k)_{k\geq 0}$ defined in a compact way as follows:

$$
\hat{u}^k = \frac{1}{S_k^\theta}\sum_{i=0}^{k-1}\theta_i u^i, \quad \text{where} \quad S_k^\theta = \sum_{i=0}^{k-1}\theta_k \quad \text{and} \quad u^i = u_\mu(x^i). \tag{10}
$$

Notice that the weights θ_k are either constant, i.e. $\theta_k = 1$ for all $k \geq 0$, or updated as $\theta_0 = 0$, $\theta_1 = 1$ and $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$.

3.2.1 Inexact gradient augmented Lagrangian method

We now analyze the overall complexity of the classical augmented Lagrangian method in terms of projections onto the cone $\mathcal{K}$ and simple feasible set U, under various assumptions on the objective function f. A direct consequence of Theorem 3.2 and Theorem 2.3 is the following iteration complexity (in terms of outer iterations) of the inexact gradient augmented Lagrangian method.

COROLLARY 3.3 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \delta > 0$ and $(x^k)_{k\geq 0}$ be the sequence generated by Algorithm **ICFG** $(d_\mu^{ag}, 0, 3\delta, 2L_d)$ with $\theta_k = 1$ for all $k \geq 0$ and $L_d = \frac{1}{\mu}$. Define the average sequence $(\hat{x}^k)_{k\geq 1}$ by $\hat{x}^k = \frac{1}{k}\sum_{i=0}^{k-1} x^{i+1}$. Then, we have the following convergence estimate on dual suboptimality:*

$$f^* - d_\mu^{ag}(\hat{x}^k) \leq \frac{L_d R_d^2}{k} + 3\delta.$$

Note that the above convergence rate is linked only to the number of the outer iterations and omits the complexity of solving the inner subproblem at step 1 of **ICFG**. Before estimating the total complexity of the process containing the inner and outer levels, we provide convergence rates for the primal infeasibility and suboptimality.

THEOREM 3.4 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \delta > 0$ and $(x^k)_{k\geq 0}$ be the sequence generated by Algorithm **ICFG**$(d_\mu^{ag}, 0, 3\delta, 2L_d)$ with $\theta_k = 1$ for all $k \geq 0$ and $L_d = \frac{1}{\mu}$. Let $u^i = u_\mu(x^i)$ be such that $\mathcal{L}_\mu^{ag}(u^i, x^i) - d_\mu^{ag}(x^i) \leq \delta$ for $0 \leq i \leq k$. Then, the average primal sequence $(\hat{u}^k)_{k\geq 1}$ defined by (10) satisfies the following relations:*
(i) The primal infeasibility is bounded sublinearly as follows:

$$dist_{\mathcal{K}}(G\hat{u}^k + g) \leq \frac{4L_d R_d}{k} + \sqrt{\frac{12L_d\delta}{k}}.$$

(ii) The primal suboptimality gap is bounded by:

$$-\frac{4L_d R_d^2}{k} - R_d\sqrt{\frac{12L_d\delta}{k}} \leq f(\hat{u}^k) - f^* \leq \frac{L_d\|x^0\|^2}{k} + 3\delta.$$

Proof. In order to facilitate an easy reading of the results, we provide the proof of primal infeasibility and suboptimality bounds in Appendix A.1. ■

Note that using the above rate of convergence, one might choose an appropriate value of the smoothing parameter μ such that after a single outer iteration an ϵ−optimal point is obtained. Thus, convergence rates in terms of outer iterations are not relevant in this case and it is natural to analyze the computational complexity of the Algorithm **ICFG**, by taking into account also the complexity of solving the inner subproblems. Therefore, we need to also count the number of fast gradient steps, which includes projections onto U and $\mathcal{K}$, matrix-vector products Gu and $G^T x$, or gradient computations $\nabla f(u)$, performed in order to attain the required inner accuracy, at a given outer iteration. Since, in the literature this is usually measured in terms of projections onto U and $\mathcal{K}$ (see e.g. [1, 8, 10]), we also use this measure of computational complexity. We further analyze the necessary number of inner projections that the inexact gradient Lagrangian method has to perform at each outer iteration. A well-known fact that we use further is

that the function $u \mapsto \text{dist}_{\mathcal{K}}(Gu+g)^2$ has Lipschitz continuous gradient with constant $\|G\|^2$ [8]. Using this observation, depending of the assumptions on the function f, we have the following inner iteration complexities for solving approximately the inner problem $\min_{u \in U} \mathcal{L}_\mu^{\text{ag}}(u,x)$ for a given x:

(i) If the function f is simple, then Algorithm $\mathbf{ICFG}(\frac{\mu}{2}\text{dist}_{\mathcal{K}}(G\cdot+g+\frac{1}{\mu}x)^2, f, 0, \mu\|G\|^2)$ returns a primal point $u_\mu(x) \in U$ such that $\mathcal{L}_\mu^{\text{ag}}(u_\mu(x),x) - d_\mu^{\text{ag}}(x) \le \delta$ after:

$$N_\delta^{\text{in}} = \left\lceil \|G\| D_U \sqrt{\frac{2\mu}{\delta}} \right\rceil$$

projections onto the primal simple feasible set $\mathcal{K} \times U$.
(ii) If the function f is not simple, but its gradient ∇f is Lipschitz continuous with constant $L_f > 0$, then the Algorithm $\mathbf{ICFG}(\mathcal{L}_\mu^{\text{ag}}(\cdot,x), 0, 0, L_f + \mu\|G\|^2)$ returns a primal point $u_\mu(x) \in U$ such that $\mathcal{L}_\mu^{\text{ag}}(u_\mu(x),x) - d_\mu^{\text{ag}}(x) \le \delta$ after:

$$N_\delta^{\text{in}} = \left\lceil D_U \sqrt{\frac{2(L_f + \mu\|G\|^2)}{\delta}} \right\rceil \tag{11}$$

projections onto the primal simple feasible set $\mathcal{K} \times U$.

Note that if we take $L_f = 0$ in the iteration complexity (11), we recover the convergence rate for the case when f is simple function. Therefore, for a uniform complexity analysis, we provide in the following result an upper bound on the total number of projections (for an optimal smoothing parameter μ) performed by the Algorithm **ICFG**, which is dependent on L_f in the following sense: with some abuse of notation for f simple function we make the convention that $L_f = 0$, and thus we obtain the computational complexity for simple functions; otherwise, if we consider $L_f > 0$, then we recover the overall complexity for the case when ∇f is L_f-Lipschitz continuous. Moreover, we assume for simplicity that $x^0 = 0$.

Theorem 3.5 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \epsilon, \delta > 0$ and $(x^k)_{k\ge 0}$ be generated by Algorithm **ICFG**$(d_\mu^{ag}, 0, 3\delta, 2L_d)$ with $\theta_k = 1$ for all $k \ge 0$. Assume that at each outer iteration k, Algorithm **ICFG**$(\frac{\mu}{2}\,\text{dist}_{\mathcal{K}}(G\cdot+g+\frac{1}{\mu}x^k)^2, f, 0, \mu\|G\|^2)$ (if f is simple and with some abuse of notation we make the convention that $L_f = 0$) or Algorithm **ICFG**$(\mathcal{L}_\mu^{ag}(\cdot,x^k), 0, 0, L_f + \mu\|G\|^2)$ (if ∇f is $L_f > 0$ Lipschitz continuous) is called to solve the inner problem and obtain a primal approximate solution $u^k = u_\mu(x^k)$ such that $\mathcal{L}_\mu^{ag}(u^k,x^k) - d_\mu^{ag}(x^k) \le \delta$. Then, by setting the optimal smoothing parameter:*

$$\mu = \max\left\{\frac{16R_d^2}{\epsilon}, \frac{L_f}{\|G\|^2}\right\} \qquad \textit{and} \qquad \delta = \frac{\epsilon}{3} \tag{12}$$

the average primal point $\hat{u}^k$ defined by (10) is $\epsilon-$optimal after a total number of

$$k = \left\lceil \sqrt{\frac{24 L_f D_U^2}{\epsilon}} + \frac{6\|G\| D_U R_d}{\epsilon} \right\rceil$$

projections onto the primal simple feasible set $\mathcal{K} \times U$.

Proof. Using the inner accuracy from (12) into Theorem 3.4, then the outer iteration complexity of the augmented Lagrangian method is given by:

$$N_\epsilon^{\text{out}} = \left\lceil \frac{16 L_d R_d^2}{\epsilon} \right\rceil = \left\lceil \frac{16 R_d^2}{\mu\epsilon} \right\rceil. \tag{13}$$

Based on the general inner complexity (11), we are able to tackle both cases: when f is simple or ∇f is $L_f > 0$ Lipschitz continuous. Minimizing the upper bound on the product $N_\epsilon^{\text{out}} N_\delta^{\text{in}}$ over positive parameters μ we get that the value of μ given in (12) is optimal up to a constant w.r.t. the total complexity. Combining (12), i.e. $\mu = \max\left\{\frac{16R_d^2}{\epsilon}, \frac{L_f}{\|G\|^2}\right\}$, with (11), we obtain the following bound on the overall complexity:

$$\begin{aligned} N_\epsilon^{\text{out}} N_\delta^{\text{in}} &\leq \left\lceil \frac{16 R_{\text{d}}^2}{\mu\epsilon} \right\rceil \left(D_U \sqrt{\frac{6(L_f + \mu\|G\|^2)}{\epsilon}} + 1 \right) \\ &\leq \left(\frac{16 R_{\text{d}}^2}{\mu\epsilon} + 1 \right) 2 D_U \sqrt{\frac{6(L_f + \mu\|G\|^2)}{\epsilon}} \\ &\overset{(12)}{=} 2 D_U \sqrt{\frac{6L_f}{\epsilon} + \frac{6\|G\|^2 R_{\text{d}}^2}{\epsilon^2}} + 1 \leq \sqrt{\frac{24 L_f D_U}{\epsilon}} + \frac{6 D_U \|G\| R_{\text{d}}}{\epsilon} + 1. \end{aligned}$$

Note that if we take $L_f = 0$ we obtain an upper bound on the overall complexity of Algorithm **ICFG** for the case when f is a simple function. ■

Remark 1 It is interesting to observe that choosing the smoothing parameter $\mu \geq \frac{16R_{\text{d}}^2}{\epsilon}$, the estimate (13) leads to the fact that the inexact gradient augmented Lagrangian method terminates after solving only once the inner subproblem. In other words, if an upper bound on R_{d} is known, then starting from an arbitrary dual initial point $x^0 \in \mathbb{R}^m$, it is sufficient to compute $u_\mu(x^0) \in U$ satisfying $\mathcal{L}_\mu^{\text{ag}}(u_\mu(x^0), x^0) - d(x^0) \leq \epsilon$ to obtain an $\epsilon-$optimal solution of (1). In particular, if $x^0 = 0$, then the inner subproblem has the form: $\min_{u\in U} f(u) + \frac{\mu}{2}\text{dist}_{\mathcal{K}}(Gu+g)^2$, which can be seen as a differentiable penalty problem. We conclude that, in the case of known R_d, the gradient augmented Lagrangian method is similar to the quadratic penalty method. ■

3.2.2 *Inexact fast gradient augmented Lagrangian method*

We further incorporate Nesterov accelerated step into the classical augmented Lagrangian method, i.e. we apply the inexact fast gradient method on the augmented dual problem. We analyze the overall complexity of the inexact fast gradient augmented Lagrangian method, under the Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1. Using the inexact oracle relation (8) and Theorem 2.3 we immediately obtain the following iteration complexity (in terms of outer iterations) of the fast gradient method:

Corollary 3.6 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \delta > 0$, and $(x^k, y^k)_{k\geq 0}$ be the sequences generated by Algorithm **ICFG**$(d_\mu^{ag}, 0, 3\delta, 2L_d)$ with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$. Then, we have the following estimate on dual suboptimality:*

$$f^* - d_\mu^{ag}(x^k) \leq \frac{4L_d R_d^2}{(k+1)^2} + 3k\delta.$$

Note that the above convergence rate is linked only to the number of the outer iterations and omits the complexity of solving the inner subproblem at step 1. Before estimating the total complexity of the process containing the inner and outer levels, we provide convergence rates for the primal infeasibility and suboptimality.

THEOREM 3.7 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \delta > 0$ and $(x^k, y^k)_{k\geq 0}$ be the sequences generated by Algorithm* ***ICFG***$(d_\mu^{ag}, 0, 3\delta, 2L_d)$ *with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$. Let $u^i = u_\mu(x^i)$ be such that $\mathcal{L}_\mu^{ag}(u^i, x^i) - d_\mu^{ag}(x^i) \leq \delta$ for $0 \leq i \leq k$. Then, the average primal sequence $(\hat{u}^k)_{k\geq 1}$ defined by* (10) *satisfies:*
(i) The primal infeasibility is bounded sublinearly as follows:

$$dist_{\mathcal{K}}\left(G\hat{u}^k + g\right) \leq \frac{8L_d R_d}{k^2} + 8\sqrt{\frac{3L_d\delta}{k}}.$$

(ii) The primal suboptimality gap is bounded by:

$$-\frac{8L_d R_d^2}{k^2} - 8R_d\sqrt{\frac{3L_d\delta}{k}} \leq f(\hat{u}^k) - f^* \leq \frac{8L_d\|x^0\|^2}{k^2} + 3k\delta.$$

Proof. We provide the proof of the primal infeasibility and suboptimality gap bounds in the Appendix A.2. ■

The necessary number of inner iterations that the inexact fast gradient augmented Lagrangian method has to perform at each outer iteration is given by (11). As in the previous section, in the following result we provide the total number of projections performed by Algorithm **ICFG**, for simple objective functions (i.e. we make the convention that $L_f = 0$) and objective functions with Lipschitz continuous gradient (i.e. we have $L_f > 0$). Moreover, we assume for simplicity that $x^0 = 0$, $R_{\rm d} > 1$ and ϵ sufficiently small.

THEOREM 3.8 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \epsilon, \delta > 0$ and $(x^k, y^k)_{k\geq 0}$ be generated by Algorithm* ***ICFG***$(d_\mu^{ag}, 0, 3\delta, 2L_d)$ *with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for $k \geq 1$. Assume that at each outer iteration k, Algorithm* ***ICFG***$(\frac{\mu}{2} dist_{\mathcal{K}}(G\cdot + g + \frac{1}{\mu}x^k)^2, f, 0, \mu\|G\|^2)$ *(if f is simple and we make the convention that $L_f = 0$) or Algorithm* ***ICFG***$(\mathcal{L}_\mu^{ag}(\cdot, x^k), 0, 0, L_f + \mu\|G\|^2)$ *(if ∇f is $L_f > 0$ Lipschitz continuous) is called to obtain an approximate solution of the inner problem $u^k = u_\mu(x^k)$ such that $\mathcal{L}_\mu^{ag}(u^k, x^k) - d_\mu^{ag}(x^k) \leq \delta$. Then, by setting the optimal smoothing parameter:*

$$\mu = \frac{16R_d^2}{\epsilon} \quad \text{and} \quad \delta = \frac{\epsilon}{24} \tag{14}$$

the average primal point $\hat{u}^k$ defined by (10) *is $\epsilon-$optimal after a total number of*

$$k = \left\lceil \frac{14L_f^{1/2}D_U}{\epsilon^{1/2}} + \frac{56R_d\|G\|D_U}{\epsilon} \right\rceil$$

projections onto the primal simple feasible set $\mathcal{K} \times U$.

Proof. First, we observe that if $R_{\rm d} > 1$, then from Theorem 3.7 the number of the outer

iterations N_ϵ^{out} satisfies:

$$N_\epsilon^{\text{out}} = \left\lceil 4R_{\text{d}} \left(\frac{L_{\text{d}}}{\epsilon}\right)^{1/2} \right\rceil = \left\lceil 4R_{\text{d}} \left(\frac{1}{\mu\epsilon}\right)^{1/2} \right\rceil.$$

for any $\mu > 0$, and by forcing both terms in Theorem 3.7 (ii) to have lower magnitudes than ϵ, then the inner accuracy δ satisfies:

$$\delta \overset{\text{Th. } 3.7(ii)}{\leq} \min\left\{\frac{\epsilon}{N_\epsilon^{\text{out}}}, \frac{\epsilon^2 N_\epsilon^{\text{out}}}{384 R_{\text{d}}^2 L_d}\right\} = \frac{\epsilon^2 N_\epsilon^{\text{out}}}{384 R_{\text{d}}^2 L_d}.$$

If one chooses $\delta = \frac{\epsilon^2 N_\epsilon^{\text{out}}}{384 R_{\text{d}}^2 L_d}$, then this inequality implies that:

$$N_\delta^{\text{in}} \leq 2D_U \sqrt{\frac{2(L_f + \mu\|G\|^2)}{\delta}} \leq 28 D_U \sqrt{\frac{(L_f + \mu\|G\|^2) R_{\text{d}}}{\mu^{1/2}\epsilon^{3/2}}}.$$

For simplicity let μ satisfy $\mu \geq \frac{L_f}{\|G\|^2}$. Then, we obtain in this case the following computational complexity:

$$\begin{aligned} N_\delta^{\text{in}} N_\epsilon^{\text{out}} &\leq \frac{42\|G\| D_U R_{\text{d}}^{1/2} \mu^{1/4}}{\epsilon^{3/4}} \left[\frac{4R_{\text{d}}}{(\mu\epsilon)^{1/2}} + 1\right] \\ &= \frac{168\|G\| D_U R_{\text{d}}^{3/2}}{\mu^{1/4}\epsilon^{5/4}} + \frac{42\|G\| D_U R_{\text{d}}^{1/2} \mu^{1/4}}{\epsilon^{3/4}}. \end{aligned}$$

Minimizing over the set $\{\mu \in \mathbb{R} \mid \mu \geq \frac{L_f}{\|G\|^2}\}$, we obtain that the best complexity is attained for $\mu = \max\left\{\frac{L_f}{\|G\|^2}, \frac{16R_{\text{d}}^2}{\epsilon}\right\}$. For a sufficiently small ϵ, the parameter μ becomes $\mu = \frac{16R_{\text{d}}^2}{\epsilon}$, which implies $N_\epsilon^{\text{out}} = 1$ and further leads to: $\delta = \frac{\epsilon}{24}$. Since $N_\epsilon^{\text{out}} = 1$, under the above choice the total number of projections onto $\mathcal{K} \times U$ required for attaining an $\epsilon-$optimal point is given by:

$$N_\epsilon^{\text{out}} N_\delta^{\text{in}} = N_\delta^{\text{in}} \leq \sqrt{\frac{8D_U^2(L_f + \mu\|G\|^2)}{\delta}} \leq \frac{14 L_f^{1/2} D_U}{\epsilon^{1/2}} + \frac{56\|G\| D_U R_{\text{d}}}{\epsilon},$$

which proves our result. Note that if we make the convention that $L_f = 0$, then we get the overall complexity for the case when f is convex and simple function. ■

It can be observed that, for an optimal choice of the smoothing parameter μ, the inexact fast gradient augmented Lagrangian method has the same computational complexity as the classical inexact gradient augmented Lagrangian method, i.e. $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ total projections onto simple set $\mathcal{K} \times U$. However, we will show the superiority of the fast variant in Section 5, when we analyze the complexity of first order augmented Lagrangian methods for attaining the optimality criteria introduced in [10].

3.3 *Adaptive inexact augmented Lagrangian method*

We have previously seen that, in the optimal case, both classical and fast augmented Lagrangian methods are dependent on the constant $\|x^*\|$ via R_d, which in general is unknown a priori. Therefore, in this section we introduce implementable variants of previous first order augmented Lagrangian methods, which approximate $\|x^*\|$ at each iteration, but maintain the same optimal computational complexities with those given in the previous theorems (up to a logarithmic factor). First, we observe that in the optimal case (when $\|x^*\|$ is known), both classical and fast augmented Lagrangian methods perform a single outer iteration in order to attain an $\epsilon-$optimal point. Therefore, we can intuitively apply a search procedure which finds an upper bound on $\|x^*\|$ in logarithmic number of steps, by performing a single outer iteration and restarting the augmented Lagrangian method. It is important to observe that this restarting strategy leads to an identical scheme for both classical and fast augmented Lagrangian methods. Throughout this section, we assume that the gradient ∇f is Lipschitz continuous with constant $L_f > 0$ (when f is simple, with a similar reasoning as given below, we can obtain the same computational complexity results).

Algorithm A-IAL (μ_0, ϵ)
Initialize $x^0 \in \mathbb{R}^n$. For $k \geq 0$ do:

(1) Compute $u^k \in U$ such that $\mathcal{L}_\mu^{\text{ag}}(u^k, x^k) - d_\mu^{\text{ag}}(x^k) \leq \frac{\epsilon}{3}$
(2) Update: $x^{k+1} = x^k + \mu_k \nabla_x \mathcal{L}_\mu^{\text{ag}}(u^k, x^k)$
(3) If $\text{dist}_\mathcal{K}(Gu^k + g) \leq \epsilon$, then **STOP** and **return** u^k, otherwise, $k = k+1$, $\mu_{k+1} = 2\mu_k$ and go to step 1.

This adaptive scheme is equivalent with the classical augmented Lagrangian method but with increasing smoothing parameter. Further, we present the computational complexity of this algorithm in the last primal point u^k and compare it with the previous results.

Theorem 3.9 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\epsilon, \mu_0 > 0$ and $(x^k)_{k\geq 0}$ be the sequence generated by Algorithm **A-IAL**(μ_0, ϵ). Assume that at each outer iteration $k \geq 0$, the Algorithm **ICFG**$(\mathcal{L}_\mu^{ag}(\cdot, x^k), 0, 0, L_f + \mu\|G\|^2)$ is called to obtain u^k such that $\mathcal{L}_\mu^{ag}(u^k, x^k) - d_\mu^{ag}(x^k) \leq \frac{\epsilon}{3}$. Then, after a total number of:*

$$\left\lceil \log_2 \left(\max\left\{ \frac{16R_d^2}{\mu_0 \epsilon}, \frac{L_f}{\mu_0 \|G\|^2} \right\} \right) \right\rceil \left[\left(\frac{6L_f D_U^2}{\epsilon} \right)^{1/2} + 1 \right] + \frac{16\|G\|R_d D_U}{\epsilon} + \frac{4L_f^{1/2} D_U}{\epsilon^{1/2}}$$

projections onto the simple set $\mathcal{K} \times U$, the last primal point u^k satisfies

$$-\epsilon\|x^*\| \leq f(u^k) - f^* \leq \epsilon, \qquad \text{dist}_\mathcal{K}(Gu^k + g) \leq \epsilon. \tag{15}$$

Proof. From Theorem 3.5, it can be seen that the inexact gradient augmented Lagrangian method performs a single outer iteration if the optimal smoothing parameter $\mu^* = \max\left\{ \frac{16R_d^2}{\epsilon}, \frac{L_f}{\|G\|^2} \right\}$ is chosen. Therefore, by iteratively doubling an arbitrary initial value μ_0 of the smoothing parameter, we attain μ^* after:

$$N_\epsilon^{\text{out}} = \left\lceil \log_2 \left(\frac{\mu^*}{\mu_0} \right) \right\rceil$$

iterations. If the optimal value μ^* is attained, a single **A-IAL** iteration would be sufficient to obtain an $\epsilon-$optimal primal point. However, since we do not know in advance f^*, we check only the feasibility criterion and stop if a prespecified accuracy is reached. This stopping criterion ensures that the final point u^k, when the algorithm stops, satisfies:

$$-\|x^*\|\epsilon \leq f(u^k) - f^* \leq \epsilon, \qquad \text{dist}_{\mathcal{K}}(Gu^k + g) \leq \epsilon.$$

From (11) it can be seen that the maximal number of projections performed by the Algorithm **A-IAL**, in order to ensure the above set of criteria, is given by:

$$\begin{aligned}
\sum_{k=1}^{N_\epsilon^{\text{out}}} N_{\epsilon,k}^{\text{in}} &= \sum_{k=1}^{N_\epsilon^{\text{out}}} \left\lceil \sqrt{\frac{6(L_f + \mu_k\|G\|^2)D_U^2}{\epsilon}} \right\rceil \\
&\leq N_\epsilon^{\text{out}} + \sum_{k=1}^{N_\epsilon^{\text{out}}} \left[\left(\frac{6L_f D_U^2}{\epsilon}\right)^{1/2} + 2^{k/2}\left(\frac{(6\mu_0)^{1/2}\|G\|D_U}{\epsilon^{1/2}}\right)\right] \\
&\leq N_\epsilon^{\text{out}} \left[\left(\frac{6L_f D_U^2}{\epsilon}\right)^{1/2} + 1\right] + 2\left(\frac{\mu^*}{\mu_0}\right)^{1/2}\left(\frac{(3\mu_0)^{1/2}\|G\|D_U}{\epsilon^{1/2}}\right) \\
&\leq N_\epsilon^{\text{out}} \left[\left(\frac{6L_f D_U^2}{\epsilon}\right)^{1/2} + 1\right] + \left(\frac{16R_d^2}{\epsilon} + \frac{L_f}{\|G\|^2}\right)^{1/2}\left(\frac{4\|G\|D_U}{\epsilon^{1/2}}\right) \\
&\leq N_\epsilon^{\text{out}} \left[\left(\frac{6L_f D_U^2}{\epsilon}\right)^{1/2} + 1\right] + \frac{16\|G\|R_d D_U}{\epsilon} + \frac{4L_f^{1/2} D_U}{\epsilon^{1/2}},
\end{aligned}$$

which proves our statement. ∎

The above result establishes that the Algorithm **A-IAL** performs $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ total projections onto simple set $\mathcal{K} \times U$ in order to obtain a primal point satisfying (15). Note that the order of the computational complexity is the same for the Algorithm **A-IAL** and for the inexact gradient augmented Lagrangian method. However, this adaptive scheme **A-IAL** has the advantage that it is implementable, i.e. the stopping criterion can be checked and the parameters of the method are computable.

3.4 *Inexact first order Lagrangian method for a modified Nesterov smoothing*

The smoothing strategy presented in the previous sections is equivalent with the application of the classical Moreau smoothing technique on the entire dual function d. Unlike this classical approach, we take in this section a new different path: we make use of the new separable structure of the dual function (6) and we only smooth the *Lagrangian* part d_U of the dual function and keep the nonsmooth part $d_{\mathcal{K}}$ unchanged. Based on (6), we introduce the following smooth approximation of d_U:

$$d_{U,\mu}(x) = \min_{u \in U} \mathcal{L}_\mu(u,x), \qquad \text{where} \quad \mathcal{L}_\mu(u,x) = f(u) + \langle x, Gu + g\rangle + \mu\, p_U(u),$$

where $p_U(u)$ is a simple prox-function, continuous and strongly convex on U. Denote $u_0 = \arg\min_{u\in U} p_U(u)$ and assume without loss of generality that $p_U(u_0) = 0$ and its strong convexity parameter is 1. Then, we have $p_U(u) \geq 1/2\|u - u_0\|^2 \ \forall u \in U$. One

typical example satisfying these assumptions is $p_U(u) = 1/2\|u\|^2$. The function $d_{U,\mu}$ has Lipschitz continuous gradient [19]:

$$\|\nabla d_{U,\mu}(x) - \nabla d_{U,\mu}(y)\| \leq L_{\mathrm{d}}\|x-y\| \quad \forall x, y \in \mathbb{R}^m, \quad \text{with constant} \quad L_{\mathrm{d}} = \|G\|^2/\mu.$$

First, note that if $\mu = 0$, then we recover the classical Lagrangian and dual functions. Secondly, the gradient of $d_{U,\mu}$ satisfies:

$$\nabla d_{U,\mu}(x) = Gu^*_\mu(x) + g, \qquad \text{where} \qquad u^*_\mu(x) \in \arg\min_{u\in U} \mathcal{L}_\mu(u, x).$$

Moreover, we use in the sequel the following notation for the partial gradient of $\mathcal{L}_\mu$:

$$\nabla_x \mathcal{L}_\mu(u, x) = Gu + g.$$

The smoothed dual function $d_{U,\mu}$ leads to a novel smooth approximation of the composite dual function d, that we aim to maximize using fast gradient method:

$$f^*_\mu = \max_{x\in\mathbb{R}^m} \; d_\mu(x) \quad \left(= d_{U,\mu}(x) + d_{\mathcal{K}}(x)\right).$$

We denote with $X^*_\mu = \arg\max_{x\in\mathbb{R}^m} \; d_\mu(x)$ the optimal solution set of the smoothed dual problem and x^*_μ an optimal point. It is important to note that, in many cases, $u^*_\mu(x)$ cannot be computed exactly, but within a pre-specified accuracy, which leads us to the inexact framework introduced in the previous section. Thus, in the rest of the section we define $u_\mu(x) \in U$ the inexact solution of the inner problem satisfying:

$$0 \leq \mathcal{L}_\mu(u_\mu(x), x) - d_{U,\mu}(x) \leq \delta. \tag{16}$$

Then, we can derive a first order inexact oracle for the smoothed dual function $d_{U,\mu}$:

THEOREM 3.10 *Let $\mu, \delta > 0$, then we have the following first order inexact $(3\delta, 2L_d)$-oracle for the smoothed dual function $d_{U,\mu}$:*

$$0 \leq \mathcal{L}_\mu(u_\mu(y), y) + \langle \nabla_x \mathcal{L}_\mu(u_\mu(y), y), x-y\rangle - d_{U,\mu}(x) \leq \frac{2L_d}{2}\|x-y\|^2 + 3\delta \tag{17}$$

for all $x, y \in \mathbb{R}^m$, where $u_\mu(y) \in U$ satisfies (16) and $L_d = \frac{\|G\|^2}{\mu}$.

Proof. The proof is similar with the one given in Theorem 3.2 and thus we omit it. ■

The relation (17) implies that the smoothed dual function $d_{U,\mu}$ is equipped with a $(3\delta, 2L_{\mathrm{d}})$-oracle, i.e. $\phi_{\delta,L}(x) = \mathcal{L}_\mu(u_\mu(x), x)$ and $\nabla\phi_{\delta,L}(x) = \nabla_x\mathcal{L}_\mu(u_\mu(x), x) = Gu_\mu(x) + g$. We notice that there are some previous results on the application of Nesterov smoothing technique for solving the dual of linear equality constrained convex problems [3, 7, 15, 23, 24], but these algorithms require exact solution of the inner subproblem and more conservative convergence estimates are derived. Further, we estimate the rate of convergence of Algorithm **ICFG** on the modified Nesterov smoothing of the dual function. First, let us redefine the following finite quantity:

$$R_{\mathrm{d}} = \max_{\mu\in\mathcal{C}} \min_{x^*_\mu \in X^*_\mu} \|x^0 - x^*_\mu\| < \infty,$$

where $\mathcal{C}$ is a compact set in $\mathbb{R}_+$. From [17][Lemma 1] it follows immediately that such an R_{d} is always finite for $\mathcal{C} = [0,\ c]$, with $0 < c < \infty$, provided that a Slater vector exists. Note that we can also bound $\min\limits_{x^* \in X^*} \|x^0 - x^*\| \leq R_{\mathrm{d}}$. Using Theorem 2.3, we get the following estimate on dual suboptimality:

Corollary 3.11 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \delta > 0$ and $(x^k, y^k)_{k \geq 0}$ be the sequences generated by Algorithm **ICFG**$(d_{U,\mu}, d_{\mathcal{K}}, 3\delta, 2L_d)$, with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for $k \geq 1$. Then, we have the following estimate on dual suboptimality:*

$$f_\mu^* - d_\mu(x^k) \leq \frac{4L_d R_d^2}{(k+1)^2} + 3k\delta.$$

We further estimate the rate of convergence of the average primal sequence generated by **ICFG** on the modified Nesterov smoothing of the dual function. For simplicity of the exposition, we assume further that $x^0 = 0,\ u^0 = 0,\ R_{\mathrm{d}} \geq 1,\ \|G\| > 1,\ D_U > 1$ and $\epsilon < 1$. However, in the case when one of these conditions does not hold, then there is no change in the order of our results, but slight differences in constants. Using these simplifications, we obtain the following outer iteration complexity for **ICFG**.

Theorem 3.12 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \delta > 0$ and $(x^k, y^k)_{k \geq 0}$ be the sequence generated by Algorithm **ICFG**$(d_{U,\mu}, d_{\mathcal{K}}, 3\delta, 2L_d)$ with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$. Define $u^i = u_\mu(x^i)$ such that $\mathcal{L}_\mu(u^i, x^i) - d_{U,\mu}(x^i) \leq \delta$. For any fixed number of outer iterations $K \geq 1$, if we set $\mu(K) = \frac{2^{3/2}\|G\| R_d}{D_U K}$, the average primal point $\hat{u}^K$ defined by (10) satisfies:*
(i) The primal infeasibility is bounded sublinearly as follows:

$$dist_{\mathcal{K}}(G\hat{u}^K + g) \leq \frac{2^{3/2}\|G\| D_U}{K} + 2\left(\frac{\|G\| D_U \delta}{R_d}\right)^{1/2}. \tag{18}$$

(ii) The primal suboptimality gap is bounded sublinearly by:

$$-\frac{2^{3/2}\|G\| D_U R_d}{K} - 2\left(\|G\| D_U R_d \delta\right)^{1/2} \leq f(\hat{u}^K) - f^* \leq \frac{2^{3/2}\|G\| R_d D_U}{K} + 3K\delta. \tag{19}$$

Proof. This proof is similar to the one given in Appendix A.2. However, it is also given in the companion paper [14, Appendix A.3]. ∎

It is important to remark that if the functions f and p_U are simple, then by definition, the solution of the inner problem $\min_{u \in U} \mathcal{L}_\mu(u, x)$ can be found efficiently (e.g. in linear time or even in closed form). Otherwise, $\mathcal{L}_\mu(\cdot, x)$ is a composition between a function f with ∇f Lipschitz continuous with constant $L_f > 0$, and a μ-strongly convex and simple function p_U and thus, Nesterov optimal method for composite problems with a strongly convex part and a smooth part finds an approximate solution $u_\mu(x)$ for the inner problem satisfying $\mathcal{L}_\mu(u_\mu(x), x) - d_{U,\mu}(x) \leq \delta$ in [21]:

$$N_\delta^{\mathrm{in}} = \left\lceil \sqrt{\frac{L_f}{\mu}} \log\left(\frac{L_f D_U^2}{4\delta}\right) \right\rceil \tag{20}$$

projections onto the simple set U. Now, we are ready to derive the overall iteration complexity of Algorithm **ICFG** in this case:

THEOREM 3.13 *Under Assumptions 2.1 with $\sigma_f = 0$, 2.2 and 3.1, let $\mu, \delta > 0$ and $\epsilon > 0$, and the sequences $(x^k, y^k)_{k\geq 0}$ be generated by the Algorithm **ICFG**$(d_{U,\mu}, d_{\mathcal{K}}, 3\delta, 2L_d)$ with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$. Also let $N_\epsilon^{out} = \left\lceil \frac{6\|G\|D_U R_d}{\epsilon} \right\rceil$. Assume further that the primal average point $\hat{u}^k$ is given by (10), then the following assertions hold:*
(i) If the function f is simple, then by setting an optimal smoothing parameter:

$$\mu = \frac{2^{3/2}\|G\|R_d}{D_U N_\epsilon^{out}} \quad \text{and} \quad \delta = 0,$$

the primal average point $\hat{u}^k$ is $\epsilon-$optimal after

$$k = \left\lceil \frac{6\|G\|R_d D_U}{\epsilon} \right\rceil$$

projections onto the primal feasible set U and polar cone $\mathcal{K}^*$.
(ii) If the function f is not simple, but ∇f is Lipschitz continuous with $L_f > 0$ and, at each outer iteration $k \geq 1$, Nesterov optimal method for strongly convex and smooth objective functions [21] is called to obtain an approximate optimal point $u^k = u_\mu(x^k) \in U$ such that $\mathcal{L}_\mu(u^k, x^k) - d_{U,\mu}(x^k) \leq \delta$. By setting an optimal smoothing parameter:

$$\mu = \frac{2^{3/2}\|G\|R_d}{D_U N_\epsilon^{out}} \quad \text{and} \quad \delta = \frac{\epsilon}{6N_\epsilon^{out}},$$

then the average primal point $\hat{u}^k$ is $\epsilon-$optimal after at most

$$k = \left\lfloor \left(\frac{24\|G\|R_d D_U^2 L_f^{1/2}}{\epsilon^{3/2}} + \frac{12 L_f^{1/2}\|G\|D_U R_d}{\epsilon} \right) \left[\log\left(\frac{36\|G\|R_d L_f D_U^3}{\epsilon^2} \right) + 1 \right] \right\rfloor$$

projections onto the set U and $\left\lceil \frac{6\|G\|R_d D_U}{\epsilon} \right\rceil$ projections onto the cone $\mathcal{K}^*$.

Proof. By forcing both hand sides in (19) to be equal with ϵ, then we obtain:

$$N_\epsilon^{\text{out}} = \left\lceil \frac{6\|G\|D_U R_{\text{d}}}{\epsilon} \right\rceil \tag{21}$$

outer projections onto $\mathcal{K}^*$, and the inner accuracy satisfies (provided that $L_f > 0$):

$$\delta = \min\left\{ \frac{\epsilon^2}{8\|G\|D_U R_{\text{d}}}, \frac{\epsilon}{6N_\epsilon^{\text{out}}} \right\} \leq \frac{\epsilon^2}{36\|G\|D_U R_{\text{d}}}. \tag{22}$$

Considering the optimal choice of the smoothing parameter (see [14, Appendix A.3]) and taking into account the bound (21) we get:

$$\mu(N_\epsilon^{\text{out}}) = \frac{2^{3/2}\|G\|R_{\text{d}}}{D_U N_\epsilon^{\text{out}}}.$$

Further, using (22) and the smoothing parameter $\mu(N_\epsilon^{\mathrm{out}})$ in the inner complexity (20), we get the following bound on the total number of inner iterations:

$$
\begin{aligned}
N_\epsilon^{\mathrm{out}} N_\delta^{\mathrm{in}} &\leq \frac{12\|G\| D_U R_{\mathrm{d}}}{\epsilon}\left(\sqrt{\frac{3L_f D_U^2}{\epsilon}} + \sqrt{L_f}\right)\log\left(\frac{36\|G\| R_{\mathrm{d}} L_f D_U^3}{\epsilon^2}\right) \\
&\qquad + \frac{12\|G\| D_U R_{\mathrm{d}}}{\epsilon} \\
&\leq \left(\frac{24\|G\| R_{\mathrm{d}} D_U^2 L_f^{1/2}}{\epsilon^{3/2}} + \frac{12 L_f^{1/2}\|G\| D_U R_d}{\epsilon}\right)\left[\log\left(\frac{36\|G\| R_{\mathrm{d}} L_f D_U^3}{\epsilon^2}\right) + 1\right].
\end{aligned}
$$

These bounds confirm our result. ■

Remark 2 If the objective function f is strongly convex, i.e. it satisfies Assumption 2.1 with $\sigma_f > 0$, and has Lipschitz gradient of constant $L_f > 0$, then it is well known that the dual function d has Lipschitz gradient with constant $L_{\mathrm{d}} = \frac{\|G\|^2}{\sigma_f}$ [13], and therefore any smoothing technique is redundant. In this setting, using the first order inexact oracle framework from previous sections, we can easily derive overall complexity of Algorithm **ICFG** for the average primal point $\hat{u}^k$ of order $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\log(\frac{1}{\epsilon^{3/2}})\right)$ projections onto the set U and $\mathcal{O}(\frac{1}{\sqrt{\epsilon}})$ projections onto the cone $\mathcal{K}$, see [13] for more details. ■

Remark 3 If we assume that there exists a bound R_p such that $\max_{x\in\mathcal{K}^*}\|u^0 - u(x)\| \leq R_p < \infty$, then we can remove the boundedness assumption on U (i.e. Assumption 2.2 (ii)) and all the previous complexity results hold by replacing D_U with R_p. ■

In conclusion, the inexact fast gradient method for the modified Nesterov smoothing of the dual function performs the same number $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ of projections onto the cone as the previous inexact first order augmented Lagrangian methods. However, in Nesterov smoothing method for smooth objective functions the number $\mathcal{O}\left(\frac{1}{\epsilon^{3/2}}\right)$ of projections onto U is significantly larger than in the previous augmented Lagrangian smoothing methods. On the other hand, the optimal smoothing parameter μ given in the augmented Lagrangian framework cannot be fixed a priori due to its dependence on $\|x^*\|$ via R_{d} and thus we need some adaptive scheme, while the optimal choice of μ in the case of Nesterov smoothing strategy can be easily computed in the initialization phase according to Theorem 3.13.

4. First order penalty methods

The complexity analysis of primal-dual methods from Section 3 has been based on the Assumption 3.1. Also the most papers on penalty methods make the strong Assumption 3.1, that is there exists an optimal Lagrange multiplier for the primal convex problem (1) [9]. This property is usually guaranteed through a Slater type condition, which in the large-scale settings it is very difficult to check computationally or such a condition might not even hold. In this section we remove Assumption 3.1 and analyze various penalty strategies for solving the conic constrained convex optimization problem (1) without this assumption. Therefore, we now consider the conic convex problem (1) which does not necessarily admit a Lagrange multiplier that closes the duality gap. To the best of

our knowledge this is one of the first computational complexity results for first order penalty methods for conic problems when it is not necessarily assumed the existence of a Lagrange multiplier that closes the duality gap.
First, denote $f_* = \min_{u \in U} f(u)$. Given the difficulties induced by the linear conic constraints, the original problem (1) can be reformulated in this case, using a (non)differentiable penalty function, as an optimization problem with simple constraints. Therefore, for a penalty parameter $\rho > 0$, the basic penalty reformulations of problem (1) are as follows:

$$\min_{u \in U} \quad \psi_\rho(u) \quad \left(= f(u) + \frac{\rho}{2}\text{dist}_{\mathcal{K}}(Gu+g)^2\right), \tag{23}$$

$$\min_{u \in U} \quad \phi_\rho(u) \quad (= f(u) + \rho\text{dist}_{\mathcal{K}}(Gu+g)). \tag{24}$$

Depending on the context, we denote $u^*_\rho \in \arg\min_{u \in U} \psi_\rho(u)$ or $u^*_\rho \in \arg\min_{u \in U} \phi_\rho(u)$. It is well-known that both formulations have certain advantages and disadvantages. The differentiable formulation (23) features good smoothness properties, but it is regarded as an inexact penalty problem, i.e. as $\rho \to \infty$ we have $u^*_\rho \to u^* \in U^*$. On the other hand, the nondifferentiable formulation (24) lacks smoothness properties, but in the case when optimal Lagrange multipliers for (1) exist, there is a finite threshold $\rho^* > 0$ such that for any $\rho \geq \rho^*$, we have $u^*_\rho = u^* \in U^*$. We recall the convexity property of the distance:

$$\text{dist}_{\mathcal{K}}(Gu+g) \geq \text{dist}_{\mathcal{K}}(Gv+g) + \langle G^T s(v), u - v\rangle \quad \forall u, v \in \mathbb{R}^m, \tag{25}$$

where $s(v) \in \partial\text{dist}_{\mathcal{K}}(Gv+g)$ denotes a subgradient at v of function $\text{dist}_{\mathcal{K}}(G\cdot+g)$. From (25), it can be easily seen that for any $u \in \mathbb{R}^m$ such that $Gu + g \in \mathcal{K}$ results:

$$\langle s(v), Gv - Gu\rangle \geq \text{dist}_{\mathcal{K}}(Gv+g) \quad \forall v \in \mathbb{R}^m. \tag{26}$$

Further we analyze both penalty strategies combined with fast gradient method and we derive the overall complexities for them.

4.1 *Fast gradient differentiable penalty method*

If the gradient ∇f is $L_f > 0$ Lipschitz continuous, then the penalty function ψ_ρ has also Lipschitz continuous gradients with constant $L_\psi = L_f + \rho\|G\|^2$. Note that the optimality conditions of (23) are:

$$\langle \nabla f(u^*_\rho) + \rho\text{dist}_{\mathcal{K}}(Gu^*_\rho + g)G^T s(u^*_\rho), u - u^*_\rho\rangle \geq 0 \quad \forall u \in U. \tag{27}$$

Now, we state our result regarding the computational complexity of the penalty method with differentiable penalty, regarding simple objective functions (set $L_f = 0$ in the complexity estimate) or smooth objective functions with Lipschitz continous gradients (i.e. $L_f > 0$). Define $\Delta^* = f^* - f_*$.

THEOREM 4.1 *Under Assumptions 2.1 with $\sigma_f = 0$ and 2.2, let $\rho > 0, \epsilon \in (0, \Delta^*/2)$ and $(u^k, v^k)_{k \geq 0}$ be the sequence generated by the Algorithm* ***ICFG***$(\psi_\rho, 0, 0, L_\psi)$ *with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for $k \geq 1$. If the penalty parameter satisfies:*

$$\rho \geq \frac{4\Delta^*}{\epsilon^2} \tag{28}$$

and f is simple ($L_f = 0$) or ∇f Lipschitz continuous ($L_f > 0$), then after

$$k = \left\lceil \sqrt{\frac{2L_f D_U^2}{\epsilon}} + \frac{(8\Delta^*)^{1/2}\|G\|D_U}{\epsilon^{3/2}} \right\rceil$$

projections onto the simple set $\mathcal{K} \times U$, we have:

$$-\Delta^* \leq f(u^k) - f^* \leq \epsilon, \qquad dist_{\mathcal{K}}(Gu^k + g) \leq \epsilon. \tag{29}$$

Proof. First, observe that by taking $u = u^*$ in (27) and using (26), we obtain:

$$\langle \nabla f(u_\rho^*), u^* - u_\rho^* \rangle \geq \rho \text{dist}_{\mathcal{K}}(Gu_\rho^* + g)^2.$$

Taking into account that $f(u_\rho^*) \geq f_*$, then from the convexity property of f results:

$$\text{dist}_{\mathcal{K}}(Gu_\rho^* + g) \leq \sqrt{\frac{f(u^*) - f_*}{\rho}} = \sqrt{\frac{\Delta^*}{\rho}}.$$

Therefore, a sufficient condition for $\text{dist}_{\mathcal{K}}(Gu_\rho^* + g) \leq \epsilon/2$ is $\rho \geq \frac{4\Delta^*}{\epsilon^2}$. Let $\bar{u} \in U$ satisfying:

$$f(\bar{u}) + \frac{\rho}{2}\text{dist}_{\mathcal{K}}(G\bar{u} + g)^2 - f(u_\rho^*) - \frac{\rho}{2}\text{dist}_{\mathcal{K}}(Gu_\rho^* + g)^2 = \psi_\rho(\bar{u}) - \psi_\rho^* \leq \epsilon. \tag{30}$$

Using the convexity property of f, (26) and (27), then the relation (30) implies:

$$\begin{aligned}
\epsilon &\geq \frac{\rho}{2}\text{dist}_{\mathcal{K}}(G\bar{u} + g)^2 - \frac{\rho}{2}\text{dist}_{\mathcal{K}}(Gu_\rho^* + g)^2 + \langle \nabla f(u_\rho^*), \bar{u} - u_\rho^* \rangle \\
&\overset{(27)}{\geq} \frac{\rho}{2}\text{dist}_{\mathcal{K}}(G\bar{u} + g)^2 - \frac{\rho}{2}\text{dist}_{\mathcal{K}}(Gu_\rho^* + g)^2 + \rho\text{dist}_{\mathcal{K}}(Gu_\rho^* + g)\langle G^T s(u_\rho^*), u_\rho^* - \bar{u} \rangle \\
&\overset{(26)}{\geq} \frac{\rho}{2}\text{dist}_{\mathcal{K}}(G\bar{u} + g)^2 + \frac{\rho}{2}\text{dist}_{\mathcal{K}}(Gu_\rho^* + g)^2 - \\
&\quad - \rho\text{dist}_{\mathcal{K}}(Gu_\rho^* + g)\left(\text{dist}_{\mathcal{K}}(Gu_\rho^* + g) + \langle G^T s(u_\rho^*), \bar{u} - u_\rho^* \rangle\right) \\
&= \frac{\rho}{2}\left[\text{dist}_{\mathcal{K}}(Gu_\rho^* + g) - \text{dist}_{\mathcal{K}}(G\bar{u} + g)\right]^2.
\end{aligned}$$

The last relation leads to:

$$\text{dist}_{\mathcal{K}}(G\bar{u} + g) \leq \sqrt{\frac{2\epsilon}{\rho}} + \text{dist}_{\mathcal{K}}(Gu_\rho^* + g).$$

For a penalty parameter satisfying (28) and $\epsilon \leq \Delta^*/2$, we reach $\epsilon-$infeasibility:

$$\text{dist}_{\mathcal{K}}(G\bar{u} + g) \leq \frac{\epsilon^{3/2}}{\sqrt{2\Delta^*}} + \frac{\epsilon}{2} \leq \epsilon.$$

To obtain suboptimality bounds, first note that the left inequality stating $f(\bar{u}) - f(u^*) \geq -\Delta^*$ is trivial. Second, the relation (30) implies:

$$f(\bar{u}) - f(u^*) \leq f(\bar{u}) + \frac{\rho}{2}\text{dist}_{\mathcal{K}}(G\bar{u} + g)^2 - f(u^*) \leq \psi_\rho(\bar{u}) - \psi_\rho^* \leq \epsilon.$$

By choosing $\rho \geq \frac{4\Delta^*}{\epsilon^2}$ and solving the differentiable penalty problem (23) with accuracy ϵ leads to an $\epsilon-$optimal point of the original problem (1) which satisfies optimality criteria (29). For any fixed penalty parameter $\rho > 0$, Algorithm $\mathbf{ICFG}(\psi_\rho, 0, 0, L_\psi)$ generates a sequence $(u^k)_{k\geq 0}$ with the convergence rate (see Theorem 2.3):

$$\psi_\rho(u^k) - \psi_\rho^* \leq \frac{2(L_f + \rho\|G\|^2)D_U^2}{(k+1)^2}.$$

This rate of convergence implies that after

$$k = \left\lceil \sqrt{\frac{2(L_f + \rho\|G\|^2)D_U^2}{\epsilon}} \right\rceil$$

projections onto $\mathcal{K} \times U$, we get $\psi_\rho(u^k) - \psi_\rho^* \leq \epsilon$. Further, taking into account the estimation of the penalty parameter (28), we can bound the previous estimate as:

$$\left\lceil \sqrt{\frac{2L_f D_U^2}{\epsilon}} + \sqrt{\frac{8\Delta^*\|G\|^2 D_U^2}{\epsilon^3}} \right\rceil.$$

Note that the last estimate implies our result. ■

The following simple example shows the tightness of our result given in Theorem 4.1.

Example 1 *Given $p > 1$, consider the following convex problem:*

$$\min_{u\in\mathbb{R}^2} \quad f(u) \quad (:= u_2) \quad s.t. \quad |u_2|^p \leq u_1, \quad u_1 = 0,$$

where $U = \{u \in \mathbb{R}^2 | \ |u_2|^p \leq u_1\}$. Note that the feasible set contains only the trivial point $(0,0)$, and implicitly we have $u_1 \geq 0$. The Slater condition does not hold in this case. First, we show that this optimization problem does not admit a Lagrange multiplier closing the duality gap. The dual problem of the above example is given by:

$$\max_{x\in\mathbb{R}} \min_{u\in\mathbb{R}^2} \quad u_2 + xu_1 \quad s.t. \quad |u_2|^p \leq u_1.$$

Since the objective function is linear, an equivalent form of the dual problem is:

$$\max_{x\in\mathbb{R}} \min_{u} \ \pm u_1^{1/p} + xu_1.$$

Considering the case $u_2 = -u_1^{1/p}$ (for the other case we can use the same reasoning), with the implicit constraint $u_1 \geq 0$, the optimal solution u_1^* of this minimization subproblem is given by: $u_1^* = (px)^{\frac{p}{1-p}}$. Replacing this value into the cost, and taking into account that we have to keep $u_1^* \geq 0$, then we obtain the dual problem:

$$\sup_{x\geq 0} \left(\frac{1}{px}\right)^{\frac{1}{p-1}} \left(\frac{1}{p} - 1\right).$$

The dual function is negative for any $x \geq 0$, and thus we do not have a bounded Lagrange multiplier attaining the supremum. Further we estimate the value of the penalty parameter

ρ such that we get $\epsilon-$infeasibility for u^*_ρ. The quadratic penalty reformulation is given by:

$$\min_{u \in \mathbb{R}^2} \quad u_2 + \frac{\rho}{2} u_1^2 \quad s.t. \quad |u_2|^p \leq u_1.$$

Observe that the minimizer u^*_ρ of the above problem is on the boundary of the feasible set, i.e. $|u_2|^p = u_1$. Then, we get the following equivalent problem:

$$\min_{u_2 \in \mathbb{R}} \quad u_2 + \frac{\rho}{2} u_2^{2p}.$$

The optimality condition of the above problem is given by $1 + \rho p [(u^*_\rho)_2]^{2p-1} = 0$, which immediately implies:

$$(u^*_\rho)_2 = \left(-\frac{1}{p\rho}\right)^{\frac{1}{2p-1}}. \tag{31}$$

From this expression and the fact that $|(u^*_\rho)_2|^p = (u^*_\rho)_1$, it can be derived that $\epsilon-$infeasibility is attained, i.e. $|(u^*_\rho)_1| \leq \epsilon$, provided that the penalty parameter satisfies:

$$\rho \geq \frac{1}{p}\left(\frac{1}{\epsilon}\right)^{2-\frac{1}{p}} = \frac{\epsilon^{1/p}}{p}\left(\frac{1}{\epsilon}\right)^2.$$

Observing that $\frac{\epsilon^{1/p}}{p}$ is a convex function of p, the minimal value of this expression is attained for $p^* = \ln(1/\epsilon)$. Replacing this value in the above estimate, we have:

$$\rho \geq \frac{\epsilon^{\frac{1}{\ln(1/\epsilon)}}}{\ln(1/\epsilon)}\left(\frac{1}{\epsilon}\right)^2 = \frac{1}{e\ln(1/\epsilon)}\left(\frac{1}{\epsilon}\right)^2,$$

where e is the Euler constant. Therefore, for this example, the penalty parameter should satisfy $\rho = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ (up to a logarithmic factor), which confirms the tightness of our result given in Theorem 4.1. ■

4.2 Fast gradient nondifferentiable penalty method

Given the nonsmoothness feature of the penalty function ϕ_ρ, we replace the nonsmooth term $\text{dist}_\mathcal{K}(G \cdot +g)$ with a basic smooth approximation. Thus, for a given smoothing parameter $\mu > 0$, we replace the original problem with the following smooth problem:

$$\min_{u \in U} \ \phi_{\rho,\mu}(u) \qquad \left(= f(u) + \rho\sqrt{\text{dist}_\mathcal{K}(Gu+g)^2 + \mu^2}\right). \tag{32}$$

Note that if ∇f is Lipschitz continuous with constant $L_f > 0$, then $\nabla \phi_{\rho,\mu}$ is Lipschitz continuous with constant $L_\phi = L_f + \frac{\rho\|G\|}{\mu}$. We denote $u^*_\mu \in \arg\min_{u \in U} \ \phi_{\rho,\mu}(u)$ and, for simplicity, assume that $\Delta^* \geq \epsilon$ (otherwise some minor changes in constants will occur).

THEOREM 4.2 *Under Assumptions 2.1 with $\sigma_f = 0$ and 2.2, let $\mu, \rho, \epsilon > 0$ and the sequence $(u^k, v^k)_{k \geq 0}$ be generated by the Algorithm **ICFG**$(\phi_{\rho,\mu}, 0, 0, L_\phi)$ with $\theta_{k+1} =$*

$\frac{1+\sqrt{1+4\theta_k^2}}{2}$ *for all* $k \geq 1$*. If the following conditions hold:*

$$\rho = \frac{2\Delta^*}{\epsilon} + 1 \quad \textit{and} \quad \mu = \frac{\epsilon}{2}, \tag{33}$$

and f *is simple (convention* $L_f = 0$*) or* ∇f *Lipschitz continuous (*$L_f > 0$*), then after*

$$k = \left\lceil \sqrt{\frac{2L_f D_U^2}{\epsilon}} + \sqrt{\frac{12\Delta^* \|G\| D_U^2}{\epsilon^3}} \right\rceil$$

projections onto the primal simple feasible set $\mathcal{K} \times U$*, we have:*

$$-\Delta^* \leq f(u^k) - f^* \leq \epsilon, \qquad \mathit{dist}_{\mathcal{K}}(Gu^k + g) \leq \epsilon.$$

Proof. Let $\bar{u} \in U$ be an $\epsilon-$optimal point for the smoothed penalty problem (32) satisfying $\phi_{\rho,\mu}(\bar{u}) - \phi^*_{\rho,\mu} \leq \epsilon$, i.e. we have:

$$f(\bar{u}) + \rho\sqrt{\text{dist}_{\mathcal{K}}(G\bar{u}+g)^2 + \mu^2} - f(u^*_\mu) - \rho\sqrt{\text{dist}_{\mathcal{K}}(Gu^*_\mu + g)^2 + \mu^2} \leq \epsilon. \tag{34}$$

First, the relation (34) implies the following:

$$\begin{aligned} f(\bar{u}) - f^* &\leq f(\bar{u}) + \rho\sqrt{\text{dist}_{\mathcal{K}}(G\bar{u}+g)^2 + \mu^2} - f^* - \rho\mu \\ &\leq f(\bar{u}) + \rho\sqrt{\text{dist}_{\mathcal{K}}(G\bar{u}+g)^2 + \mu^2} - f(u^*_\mu) - \rho\sqrt{\text{dist}_{\mathcal{K}}(Gu^*_\mu + g)^2 + \mu^2} \leq \epsilon. \end{aligned} \tag{35}$$

Second, from (34) we have the following feasiblity relation:

$$\begin{aligned} \text{dist}_{\mathcal{K}}(G\bar{u}+g) &\leq \sqrt{\text{dist}_{\mathcal{K}}(G\bar{u}+g)^2 + \mu^2} \\ &\overset{(35)}{\leq} \frac{f(u^*_\mu) + \rho\sqrt{\text{dist}_{\mathcal{K}}(Gu^*_\mu + g)^2 + \mu^2} - f(\bar{u}) + \epsilon}{\rho} \\ &\leq \frac{f^* - f_* + \epsilon}{\rho} + \mu = \frac{\Delta^* + \epsilon}{\rho} + \mu. \end{aligned}$$

Therefore, choosing the parameters conformal to (33), any point satisfying (34) is $\epsilon-$optimal in the optimality criteria (29). Given arbitrary $\mu, \rho > 0$, the Algorithm **ICFG**$(\phi_{\rho,\mu}, 0, 0, L_\phi)$ applied on the smoothed problem (32) generates primal sequences $(u^k, v^k)_{k\geq 0}$ satisfying the following convergence rate (see Theorem 2.3):

$$\phi_{\rho,\mu}(u^k) - \phi^*_{\rho,\mu} \leq \frac{2\left(L_f + \rho\frac{\|G\|}{\mu}\right) D_U^2}{k^2}.$$

Thus, the $\epsilon-$suboptimality for problem (32) is attained after at most:

$$\left\lceil \sqrt{\frac{2L_f D_U^2}{\epsilon}} + \sqrt{\frac{2\rho\|G\| D_U^2}{\mu\epsilon}} \right\rceil$$

projections onto the set $\mathcal{K} \times U$. Taking into account the assumptions (33), we obtain the computational complexity estimate given in the theorem. ■

Remark 4 It is easy to prove that if the objective function f is strongly convex, i.e. it satisfies Assumption 2.1 with $\sigma_f > 0$, then the differentiable and nondifferentiable penalty methods from previous sections have computational complexity in the last primal point u^k of order $\mathcal{O}\left(\frac{1}{\epsilon}\log(\frac{1}{\epsilon})\right)$ projections onto the set $\mathcal{K} \times U$. ■

From previous discussion it follows that the optimal penalty parameter ρ depends on Δ^*, which in general is unknown a priori. Therefore, in the next section we introduce implementable variants of previous first order penalty methods, which approximate Δ^* at each iteration, but maintain the same optimal computational complexities with those given in the previous theorems (up to a logarithmic factor).

4.3 *Adaptive fast gradient penalty method*

In this section, regardless of the type of penalty function, we introduce an Adaptive Penalty Method (A-PM), which rely on a sequential increase of the penalty parameter ρ until a satisfactory value is attained.

Algorithm A-PM (ρ_0, ϵ, s)

1. Set $k = 0$ and choose $u_0 \in U$. If $s =$ "N" choose $\mu > 0$. For $k \geq 0$ do:
2. Apply the Algorithm **ICFG** on the (smoothed) penalty subproblem and find u^k such that:

$$\psi_{\rho_k}(u^k) - \psi^*_{\rho_k} \leq \epsilon, \text{ if } \quad s = \text{“}D\text{''};$$
$$\phi_{\rho_k}(u^k) - \phi^*_{\rho_k} \leq \epsilon, \text{ if } \quad s = \text{“}N\text{''}.$$

3. If the iterate u^k satisfies $\text{dist}_{\mathcal{K}}(Gu^k + g) \leq \epsilon$, then **STOP**. Otherwise, set $\rho_{k+1} = 2\rho_k, k = k + 1$ and go to step 2.

In the previous sections we have seen that, in the general case, when the optimal Lagrange multipliers do not necessarily exist, there is a penalty parameter $\bar{\rho}$ dependent on the type of penalty function, i.e.: $\bar{\rho} = \begin{cases} \frac{4\Delta^*}{\epsilon^2}, & \text{for smooth penalty} \\ \frac{3\Delta^*}{\epsilon}, & \text{for nonsmooth penalty} \end{cases}$, such that if $\rho_k \geq \bar{\rho}$ and $\epsilon \leq \Delta^*/2$, then u^k satisfies (29) and the algorithm stops. Further, we provide the computational complexity for Algorithm **A-PM** in the case when ∇f is Lipschitz continuous with constant $L_f > 0$. The complexity results for the case when f is simple can be derived similarly.

Theorem 4.3 *Under the assumptions of Theorem 4.1, let $\rho_0, \epsilon > 0$ and the sequence $(u^k)_{k\geq 0}$ be generated by Algorithm **A-PM**(ρ_0, ϵ, s). For nondifferentiable penalty case assume $\mu = \frac{\epsilon}{2}$. After a total number of projections onto $\mathcal{K} \times U$ given by:*

$$\begin{cases} \left\lceil N^{out}_{\epsilon} \left(\frac{4L_f D_U^2}{\epsilon}\right)^{1/2} + \frac{24(\rho_0\Delta^*)^{1/2}\|G\|D_U}{\epsilon^{3/2}} \right\rceil, & \textit{for smooth penalty} \\ \left\lceil N^{out}_{\epsilon} \left(\frac{4L_f D_U^2}{\epsilon}\right)^{1/2} + \frac{30(\rho_0\Delta^*\|G\|)^{1/2}D_U}{\epsilon^{3/2}} \right\rceil, & \textit{for nonsmooth penalty,} \end{cases}$$

where $N_\epsilon^{out} = \begin{cases} \left\lceil \log\left(\frac{4\Delta^*}{\epsilon^2\rho_0}\right)\right\rceil, & \textit{for smooth penalty} \\ \left\lceil \log\left(\frac{3\Delta^*}{\epsilon\rho_0}\right)\right\rceil, & \textit{for nonsmooth penalty} \end{cases}$, *the primal point* u^k *satisfies primal suboptimality* $f(u^k) - f^* \leq \epsilon$ *and primal infeasibility* $dist_{\mathcal{K}}(Gu^k + g) \leq \epsilon$.

Proof. The proof follows similar lines as in Theorem 3.9. It can be easily seen that, independently of the assumptions on the objective function f, Algorithm **A-PM** requires:

$$N_\epsilon^{\text{out}} = \begin{cases} \left\lceil \log\left(\frac{4\Delta^*}{\epsilon^2\rho_0}\right)\right\rceil, & \text{for smooth penalty} \\ \left\lceil \log\left(\frac{3\Delta^*}{\epsilon\rho_0}\right)\right\rceil, & \text{for nonsmooth penalty} \end{cases}$$

outer steps to attain an $\epsilon-$optimal point. Taking into account that in the nonsmooth case, we apply the classical smoothing strategy from Section 4.2, the iteration complexity for solving the inner subproblem, at outer iteration k, can be bounded by:

$$N_{\epsilon,k}^{\text{in}} = \begin{cases} \left\lceil \left(\frac{2L_f D_U^2}{\epsilon}\right)^{1/2} + \rho_k^{1/2}\left(\frac{2\|G\|D_U}{\epsilon^{1/2}}\right)\right\rceil, & \text{for smooth penalty} \\ \left\lceil \left(\frac{2L_f D_U^2}{\epsilon}\right)^{1/2} + \rho_k^{1/2}\left[\frac{2\|G\|D_U}{(\mu\epsilon)^{1/2}}\right]\right\rceil, & \text{for nonsmooth penalty,} \end{cases}$$

where $\mu > 0$ is the smoothing parameter. Knowing the maximal number of outer stages, note that the total number of the fast gradient iterations can be computed by summation of all quantities $N_{\epsilon,k}^{\text{in}}$. Observing that $\sum_{k=0}^{N_\epsilon^{\text{out}}} \rho_k^{\frac{1}{2}} \leq 6\rho_0 2^{\frac{N_\epsilon^{\text{out}}}{2}}$, then we obtain the following bound on the overall complexity:

$$\sum_{k=0}^{N_\epsilon^{\text{out}}} N_{\epsilon,k}^{\text{in}} \leq \begin{cases} N_\epsilon^{\text{out}}\left(\frac{2L_f D_U^2}{\epsilon}\right)^{1/2} + \frac{24(\rho_0\Delta^*)^{1/2}\|G\|D_U}{\epsilon^{3/2}} + 1, & \text{for smooth penalty} \\ N_\epsilon^{\text{out}}\left(\frac{2L_f D_U^2}{\epsilon}\right)^{1/2} + \frac{30(\rho_0\Delta^*)^{1/2}\|G\|D_U}{\epsilon^{3/2}} + 1, & \text{for nonsmooth penalty,} \end{cases}$$

which proves the statements of the theorem. ■

Remark 5 If we assume that there exist a bound R_p such that $\|u^0 - u_\rho^*\| \leq R_p < \infty$ for all $u_\rho^* \in \arg\min_{u\in U} \psi_\rho(u)$ (or $\phi_\rho(u)$), then we can remove the boundedness assumption on U (i.e. Assumption 2.2 (*ii*)) and all the previous complexity results hold by replacing D_U with R_p. ■

In conclusion, if we do not assume the existence of an optimal Lagrange multiplier that closes the duality gap for the cone constrained convex problem (1), the computational complexity of fast gradient penalty methods, in the worst-case, is of order $\mathcal{O}(\frac{1}{\epsilon^{3/2}})$. Moreover, these bound are tight as Example 1 shows.

5. Comparisons with previous work

We now present a brief comparison of our computational complexity results on Lagrangian and penalty methods with previous complexity results from the literature in various optimality criteria. We start comparing the computational complexity results on (fast) gradient augmented Lagrangian methods in the optimality criteria used in

this paper: $|f(u_\epsilon) - f^*| \leq \epsilon$ and $\mathrm{dist}_{\mathcal{K}}(Gu_\epsilon + g) \leq \epsilon$. Various first order augmented Lagrangian methods have been developed in e.g. [1, 10] and computational complexity estimates of order $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ have been obtained in our criteria. For example, in [1], an adaptive augmented Lagrangian method for cone constrained convex optimization models was analyzed. The authors in [1] prove that the outer complexity is of order $\mathcal{O}(\log(1/\epsilon))$ and the inner accuracy is constrained to be of order $\delta_k = \mathcal{O}\left(\frac{1}{k^2\beta^k}\right)$, where $\beta > 1$, and thus the overall complexity is similar to the estimates given in our paper (up to a logarithmic factor). However, our augmented Lagrangian algorithms can be easily implemented in practice, their parameters are easy to compute and our analysis based on the inexact oracle framework is more simple and intuitive in comparison with those given in [1, 10], opening various possibilities for extensions to more complex optimization models.

On the other hand, Lan et. al. in [10] considered the linear equality constrained case (i.e. $\mathcal{K} = \{0\}$) and used another set of $\epsilon-$optimality criteria, i.e. any $u_\epsilon \in U$ is $\epsilon-$optimal if there exists $x_\epsilon \in \mathbb{R}^m$ satisfying:

$$\nabla f(u_\epsilon) + G^T x_\epsilon \in -\mathcal{N}_U(u_\epsilon) + \mathcal{B}_\epsilon(0) \quad \text{and} \quad \|Gu_\epsilon + g\| \leq \epsilon. \tag{36}$$

In these criteria, without any regularization of the original problem, the gradient augmented Lagrangian algorithm **I-AL** introduced in [10] has computational complexity of order $\mathcal{O}\left(\epsilon^{-\frac{7}{4}}\right)$. We further show that, using our approach we obtain a suboptimal point satisfying (36), with a much better iteration complexity for the same algorithm **I-AL**. More precisely, with our analysis, Theorem 3.4 leads to the fact that **I-AL** method from [10] should perform: $N^{\mathrm{out}}_{\epsilon,1} = \left\lceil \frac{16R_d}{3\mu\epsilon} \right\rceil$ outer iterations with inner accuracy $\delta = \frac{\mu\epsilon^2}{128}$. Therefore, denoting the inner complexity $N^{\mathrm{in}}_{\delta} \leq \sqrt{\frac{2^9(L_f+\mu\|G\|^2)D_U^2}{\mu\epsilon^2}}$, the first stage of **I-AL** method of [10] requires $N^{\mathrm{out}}_{\epsilon,1} N^{\mathrm{in}}_{\delta}$ projections onto U and, on the other hand, the **Postprocessing** procedure in **I-AL** of [10] performs $\left\lceil \frac{2^{5/2}(L_f+\mu\|G\|^2)D_U}{\epsilon} \right\rceil$ projections onto U. Using these bounds, for any $\mu \geq \frac{L_f}{\|G\|^2}$, the total number of projections required by the **I-AL** method in [10] is bounded with our analysis by: $\frac{2^{10}\|G\|D_U R_d}{3\mu\epsilon^2} + \frac{2^{7/2}\mu\|G\|^2 D_U}{\epsilon}$. For an optimal complexity, we choose the smoothing parameter as $\mu = \frac{2^{13/4} R_d^{1/2}}{\|G\|^{1/2}\epsilon^{1/2}} + \frac{L_f}{\|G\|^2}$. With this choice, the **I-AL** method from [10] performs with our analysis:

$$\left\lceil \mathcal{O}\left(\frac{\|G\|^{3/2} R_d^{1/2} D_U}{\epsilon^{3/2}}\right) + \mathcal{O}\left(\frac{L_f D_U}{\epsilon}\right) \right\rceil$$

projections onto U, for attaining an ϵ-optimal point w.r.t. optimality criteria (36).

Moreover, using a straightforward modification of the first stage of the **I-AL** method by replacing the outer dual gradient method with an outer dual fast gradient method, we can obtain a fast **I-AL** method. From Theorem 3.7 we have that fast **I-AL** method performs: $N^{\mathrm{out}}_{\epsilon,2} = \left\lceil \sqrt{\frac{8R_d}{\mu\epsilon}} \right\rceil$ outer iterations with inner accuracy $\delta = \frac{\mu\epsilon^2}{128}$, to attain an $\epsilon-$optimal point satisfying (36). Using the same reasoning as in the previous case, the first stage of fast **I-AL** method requires $N^{\mathrm{out}}_{\epsilon,2} N^{\mathrm{in}}_{\delta}$ projections onto U and the **Postprocessing** procedure performs $\left\lceil \frac{2^{5/2}(L_f+\mu\|G\|^2)D_U}{\epsilon} \right\rceil$ projections onto U. Using these bounds, for any $\mu \geq \frac{L_f}{\|G\|^2}$, the total number of projections required by the fast **I-AL** method is bounded

by: $\frac{2^{15/2} R_d^{1/2} \|G\| D_U}{\mu^{1/2} \epsilon^{3/2}} + \frac{2^{7/2} \mu \|G\|^2 D_U}{\epsilon}$. In order to attain the optimal complexity, we choose the smoothing parameter as $\mu = \frac{4 R_d^{1/3}}{\epsilon^{1/3} \|G\|^{2/3}} + \frac{L_f}{\|G\|^2}$. With this choice, the fast **I-AL** method performs with our analysis:

$$\left\lceil \mathcal{O}\left(\frac{\|G\|^{4/3} R_d^{1/3} D_U}{\epsilon^{4/3}} \right) + \mathcal{O}\left(\frac{L_f D_U}{\epsilon} \right) \right\rceil$$

projections onto U. In conclusion, based on our settings we obtain computational complexities of order $\mathcal{O}(\epsilon^{-\frac{3}{2}})$ for the original **I-AL** method and of order $\mathcal{O}(\epsilon^{-\frac{4}{3}})$ for the fast **I-AL** method, which are significantly better than the estimate $\mathcal{O}(\epsilon^{-\frac{7}{4}})$ given in [10] for optimality criteria (36). Moreover, in our optimality criteria defined in Section 2, we have seen that for an optimal smoothing parameter both classical and fast augmented Lagrangian methods have the same complexity, while in optimality criteria (36) the fast **I-AL** has the best overall complexity. Finally, we can combine our approach with a regularization technique, i.e. the addition of a strongly convex term $\frac{\gamma}{2}\|u - u^0\|^2$ to the objective function, used e.g. in [10], and obtain also computational complexity (for the last primal point) of order $\mathcal{O}(\epsilon^{-1})$ in optimality criteria (36). Due to space limitations we omit these derivations.

Outer complexity estimate of order $\mathcal{O}(\frac{1}{\epsilon})$ for fast gradient Nesterov type smoothing methods were derived e.g. in [3, 15, 23]. From our previous analysis we can conclude that for an adequate choice of the parameter μ, the number of outer iterations is only one, and therefore, the outer complexity estimates are irrelevant to the total complexity of the method. Thus, we need to derive overall complexities as we do in this paper.

Finally, there are very few iteration complexity results for first order methods for convex problems that might not have a Lagrange multiplier closing the duality gap. Recently, Nesterov has proposed a specialized subgradient method for solving directly general nonsmooth convex problems with functional constraints without assuming the existence of bounded optimal Lagrange multipliers [22]. The specialized subgradient method in [22] requires $\mathcal{O}(\frac{1}{\epsilon^2})$ total subgradient computations for either the objective function or for a functional constraint. In [9] the classical quadratic penalty scheme is combined with Nesterov optimal method for solving a general conic problem, but under the strong assumption of the existence of optimal Lagrange multipliers. If the objective function is smooth, then the quadratic penalty method requires $\mathcal{O}(\frac{1}{\epsilon^2})$ projections on the simple convex set and on the cone to attain an ϵ-solution satisfying a criterion given in terms of a set of KKT conditions. On the other hand, using a regularization strategy for the original problem, the quadratic penalty method requires $\mathcal{O}(\frac{1}{\epsilon}\log\frac{1}{\epsilon})$ projections to attain ϵ-solution for the same criterion. Therefore, the assumption on the existence of an optimal Lagrange multiplier improves the iteration complexity of a quadratic penalty method from $\mathcal{O}(\frac{1}{\epsilon^{3/2}})$ (see Section 4) to $\mathcal{O}(\frac{1}{\epsilon})$ total iterations. Moreover, for this particular setting, one can guarantee that the suboptimality estimates hold in both sides with arbitrary accuracy, compared with our setting where only the right hand side can be attained arbitrarily small. In conclusion, the *price* we pay for tackling a more general conic convex problem is the additional computational effort and the fact that the function value represents a lower approximation of the optimal value.

6. Appendix

In this section we provide proofs for Theorems 3.4 and 3.7.

Appendix A.1

Proof of Theorem 3.4. We derive the sublinear estimates for primal infeasibility and primal suboptimality for the average primal point $\hat{u}^k = \frac{1}{k}\sum_{j=1}^{k} u^j$, where $u^j = u_\mu(x^j)$ and x^j is generated by Algorithm **ICFG**$(d_\mu^{\text{ag}}, 0, 3\delta, 2L_{\text{d}})$ with $\theta_k = 1$ for all $k \geq 0$. First, given the definition of x^{j+1} in Algorithm **ICFG** we get:

$$x^{j+1} = x^j + \frac{1}{2L_{\text{d}}}\nabla_x \mathcal{L}_\mu^{\text{ag}}(u^j, x^j) \quad \forall j \geq 0.$$

Subtracting x^j from both sides, adding up these inequalities for $j = 0 : k-1$, we get:

$$\left\| \frac{1}{k}\sum_{j=0}^{k-1} \nabla_x \mathcal{L}_\mu^{\text{ag}}(u^j, x^j) \right\| = \frac{2L_{\text{d}}}{k}\|x^k - x^0\|.$$

Note that $\nabla_x \mathcal{L}_\mu^{\text{ag}}(u^j, x^j) = Gu^j + g - \left[Gu^j + g + \frac{1}{\mu}x^j\right]_{\mathcal{K}}$. Using notation $z^j = \left[Gu^j + g + \frac{1}{\mu}x^j\right]_{\mathcal{K}}$, then $\frac{1}{k}\sum_{j=0}^{k-1} z^j \in \mathcal{K}$. This fact implies:

$$\text{dist}_{\mathcal{K}}(G\hat{u}^k + g) \leq \left\| \frac{1}{k}\sum_{j=0}^{k-1}(Gu^j + g) - \frac{1}{k}\sum_{j=0}^{k-1} z^j \right\| = \frac{2L_{\text{d}}}{k}\|x^k - x^0\|. \tag{37}$$

It remains to bound $\|x^k - x^0\|$. Using the iteration of **ICFG**, for $x \in \mathcal{K}^*$, we get:

$$\begin{aligned}
\|x^{k+1} - x\|^2 &= \|x^k - x\|^2 + 2\langle x^{k+1} - x^k, x^{k+1} - x\rangle - \|x^{k+1} - x^k\|^2 \\
&= \|x^k - x\|^2 + \frac{1}{L_{\text{d}}}\langle \nabla_x \mathcal{L}_\mu^{\text{ag}}(u^k, x^k), x^k - x\rangle \\
&\quad + \frac{1}{L_{\text{d}}}\left(\langle \nabla_x \mathcal{L}_\mu^{\text{ag}}(u^k, x^k), x^{k+1} - x^k\rangle - L_{\text{d}}\|x^{k+1} - x^k\|^2\right) \qquad (38) \\
&\leq \|x^k - x\|^2 + \frac{1}{L_{\text{d}}}(d_\mu^{\text{ag}}(x^{k+1}) - d_\mu^{\text{ag}}(x)) + \frac{3\delta}{L_{\text{d}}} \quad \forall k \geq 0.
\end{aligned}$$

Taking $x = x^*$ in the last inequality and using an inductive argument, then we get:

$$\|x^k - x^0\| \leq \|x^k - x^*\| + \|x^0 - x^*\| \leq 2\|x^0 - x^*\| + \sqrt{\frac{3k\delta}{L_{\text{d}}}}.$$

We substitute this bound into (37) and we get the estimate on primal infeasibility:

$$\text{dist}_{\mathcal{K}}(G\hat{u}^k + g) \leq \frac{4L_{\text{d}}R_{\text{d}}}{k} + \frac{2L_{\text{d}}}{k}\sqrt{\frac{3k\delta}{L_{\text{d}}}} = \frac{4L_{\text{d}}R_{\text{d}}}{k} + \sqrt{\frac{12L_{\text{d}}\delta}{k}}. \tag{39}$$

It remains to derive the estimates on primal suboptimality. First, we observe that for

any $u \in U$, we have $d_\mu(x) \leq f^*$ and the following identity holds:

$$\mathcal{L}_\mu^{\mathrm{ag}}(u,x) - \langle \nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u,x), x\rangle = f(u) + \frac{\mu}{2}\|\nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u,x)\|^2. \tag{40}$$

Based on the previous discussion, from (38) and (40) we derive that:

$$\begin{aligned}\|x^{k+1}-x\|^2 &\leq \|x^k-x\|^2 + \frac{1}{L_{\mathrm{d}}}\left(d_\mu(x^{k+1}) - \mathcal{L}_\mu(u^k,x^k) + \langle \nabla_x \mathcal{L}_\mu(u^k,x^k), x^k - x\rangle + 3\delta\right)\\ &\leq \|x^k-x\|^2 + \frac{1}{L_{\mathrm{d}}}\left(f^* - f(u^k) - \frac{\mu}{2}\|\nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u,x)\|^2 + 3\delta - \langle \nabla_x \mathcal{L}_\mu(u^k,x^k), x\rangle\right).\end{aligned}$$

Taking now $x = 0$, and using an inductive argument over $j = 0 : k-1$, we obtain:

$$f(\hat{u}^k) - f^* \leq \frac{L_{\mathrm{d}}\|x^0\|^2}{k} + 3\delta. \tag{41}$$

On the other hand, to bound below $f(\hat{u}^k) - f^*$ we proceed as follows:

$$\begin{aligned}f^* &= \min_{u\in U, r\in\mathcal{K}} f(u) + \langle x^*, Gu + g - r\rangle \leq f(\hat{u}^k) + \langle x^*, G\hat{u}^k + g - \left[G\hat{u}^k + g\right]_{\mathcal{K}}\rangle\\ &\leq f(\hat{u}^k) + \|x^*\|\|G\hat{u}^k + g - \left[G\hat{u}^k + g\right]_{\mathcal{K}}\| = f(\hat{u}^k) + \|x^*\|\mathrm{dist}_{\mathcal{K}}\left(G\hat{u}^k + g\right).\end{aligned} \tag{42}$$

Combining (39) with (42) and then with (41), we obtain the estimate on primal suboptimality stated in the theorem. ■

Appendix A.2

Proof of Theorem 3.7. We derive sublinear estimates for primal infeasibility and suboptimality of the average primal point $\hat{u}^k = \frac{1}{S_k^\theta}\sum_{j=0}^{k-1}\theta_j u^j$, where $u^j = u_\mu(x^j)$ and x^j generated by Algorithm **ICFG**$(d_\mu^{\mathrm{ag}}, 0, 3\delta, 2L_{\mathrm{d}})$ with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$. We observe that: $\frac{k+1}{2} \leq \theta_k \leq k$ and $S_k^\theta = \theta_{k-1}^2$. We denote $l^k = x^{k-1} + \theta_k(x^k - x^{k-1})$ and recall that the following relation has been proved in [13, 26]:

$$\theta_k^2(d_\mu^{\mathrm{ag}}(x) - d_\mu^{\mathrm{ag}}(x^k)) + \sum_{i=1}^{k-1}\theta_i\Delta(x,y^i) + L_{\mathrm{d}}\|l^k - x\|^2 \leq L_{\mathrm{d}}\|x^0 - x\|^2 + 3\sum_{i=1}^{k-1}\theta_i^2\delta, \tag{43}$$

where $\Delta(x,y) = \mathcal{L}_\mu^{\mathrm{ag}}(u_\mu(y),y) + \langle \nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u_\mu(y),y), x - y\rangle - d_\mu^{\mathrm{ag}}(x)$. Now we are ready to prove Theorem 3.7. From definition of augmented dual function d_μ^{ag}, it can be seen that $x^k = y^k + \frac{1}{2L_{\mathrm{d}}}\nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u^k, y^k)$. Multiplying by θ_k, we obtain:

$$\begin{aligned}\frac{\theta_k}{2L_{\mathrm{d}}}\nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u^k,y^k) &= \theta_k(x^k - y^k) = \theta_k(x^k - x^{k-1}) + (\theta_{k-1}-1)(x^{k-2} - x^{k-1})\\ &= \underbrace{x^{k-1} + \theta_k(x^k - x^{k-1})}_{l^k} - \underbrace{(x^{k-2} + \theta_{k-1}(x^{k-1} - x^{k-2}))}_{l^{k-1}}.\end{aligned} \tag{44}$$

Summing on the history of l^k and multiplying by $\frac{2L_\mathrm{d}}{S_k^\theta}$, we obtain:

$$\mathrm{dist}_\mathcal{K}\left(G\hat{u}^k + g\right) \leq \left\| \sum_{j=0}^{k-1} \frac{\theta_j}{S_k^\theta} \nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u^j, y^j) \right\| = \frac{2L_\mathrm{d}}{S_k^\theta}\|l^k - l^0\| \leq \frac{8L_\mathrm{d}}{k^2}\|l^k - l^0\|.$$

Since $x^* = \arg\max_{x\in\mathbb{R}^m} d_\mu^{\mathrm{ag}}(x)$, by taking $x = x^*$ in (43), we get:

$$\begin{aligned}\|l^k - x^*\| \leq \sqrt{\|x^0 - x^*\|^2 + \sum_{i=1}^{k-1} \frac{3\theta_i^2\delta}{L_\mathrm{d}}} &\leq \|x^0 - x^*\| + \sqrt{\frac{3\delta}{L_\mathrm{d}} S_k^\theta \max_{1\leq i\leq k-1} \theta_i} \\ &\leq \|x^0 - x^*\| + \sqrt{\frac{3\delta}{L_\mathrm{d}}}(k-1)^{3/2},\end{aligned}$$

for all $k \geq 0$. Thus, we can further bound the primal feasibility as follows:

$$\mathrm{dist}_\mathcal{K}(G\hat{u}^k + g) \leq \frac{8L_\mathrm{d}R_\mathrm{d}}{k^2} + 8\sqrt{\frac{3L_\mathrm{d}\delta}{k}}. \tag{45}$$

Further, we derive sublinear estimates for primal suboptimality. First, note that:

$$\begin{aligned}\Delta(x, y^k) &= \mathcal{L}_\mu^{\mathrm{ag}}(u^k, y^k) + \langle \nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u^k, y^k), x - y^k\rangle - d_\mu^{\mathrm{ag}}(x) \\ &\geq f(u^k) + \langle \nabla_x \mathcal{L}^{\mathrm{ag}}(u^k, y^k), x\rangle - d_\mu^{\mathrm{ag}}(x).\end{aligned}$$

Summing on the history and using the convexity of f, we get:

$$\begin{aligned}\sum_{i=1}^{k-1} \theta_i \Delta(x, y^i) &\geq \sum_{i=1}^{k-1} \theta_i (f(u^i) + \langle \nabla_x \mathcal{L}^{\mathrm{ag}}(u^i, y^i), x\rangle - d_\mu^{\mathrm{ag}}(x)) \\ &\geq \theta_k^2 \left(f(\hat{u}^k) + \sum_{i=1}^{k-1} \frac{\theta_i}{S_k^\theta} \langle \nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u^i, y^i), x\rangle - d_\mu^{\mathrm{ag}}(x) \right),\end{aligned} \tag{46}$$

for all $x \in \mathbb{R}^m$. Using (46) in (43), and dropping the term $L_\mathrm{d}\|l^k - x\|^2$, we have:

$$f(\hat{u}^k) + \sum_{i=1}^{k-1} \frac{\theta_i}{S_k^\theta} \langle \nabla_x \mathcal{L}_\mu^{\mathrm{ag}}(u^i, y^i), x\rangle - d_\mu^{\mathrm{ag}}(x) \overset{(46)+(43)}{\leq} \frac{L_\mathrm{d}}{\theta_{k-1}^2}\|x^0 - x\|^2 + \frac{3\sum_{i=1}^{k-1}\theta_i^2}{\theta_{k-1}^2}\delta$$

for all $x \in \mathbb{R}^m$. Given that $\frac{1}{\theta_{k-1}^2}\sum_{i=1}^{k-1}\theta_i^2 = \frac{1}{S_k^\theta}\sum_{i=1}^{k-1}\theta_i^2 \leq \max_{1\leq i\leq k-1} \theta_i \leq k-1$ and $d_\mu^{\mathrm{ag}}(x) \leq f^*$, by choosing the Lagrange multiplier $x = 0$, we further have:

$$f(\hat{u}^k) - f^* \leq f(\hat{u}^k) - d_\mu^{\mathrm{ag}}(0) \leq \frac{4L_\mathrm{d}\|x^0\|^2}{k^2} + 3k\delta. \tag{47}$$

On the other hand, we have:

$$f^* = \min_{u \in U, s \in \mathcal{K}} f(u) + \langle x^*, Gu + g - s \rangle \leq f(\hat{u}^k) + \langle x^*, G\hat{u}^k + g - \Big[Gu^k + g \Big]_{\mathcal{K}} \rangle$$
$$\stackrel{(45)}{\leq} f(\hat{x}^k) + \frac{8L_{\mathrm{d}}R_{\mathrm{d}}^2}{k^2} + 8R_{\mathrm{d}}\sqrt{\frac{3L_{\mathrm{d}}\delta}{k}}. \tag{48}$$

Finally, from (45), (47) and (48) we get the estimates on primal infeasibility and suboptimality stated in the theorem. ■

Appendix A.3

of Theorem 3.13. First, note that an analog result as in the previous Appendix holds in this case, and for clarity we state it below (see e.g. [26] for a proof):

LEMMA 6.1 *Let $\mu, \delta > 0$ and sequences $(x^k, y^k)_{k \geq 0}$ be generated by Algorithm* ***ICFG***$(d_{U,\mu}, d_{\mathcal{K}}, \delta)$ *with $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$ for all $k \geq 1$, then for any Lagrange multiplier x and iteration k we have:*

$$\theta_k^2(d_\mu(x) - d_\mu(x^k)) + \sum_{i=1}^{k-1} \theta_i \Delta(x, y^i) + L_d \|l^k - x\|^2 \leq L_d \|x^0 - x\|^2 + 3\sum_{i=1}^{k} \theta_i^2 \delta, \tag{49}$$

where we use $\Delta(x, y) = \mathcal{L}_\mu(u_\mu(y), y) + \langle \nabla_x \mathcal{L}_\mu(u_\mu(y), y), x - y \rangle - d_\mu(x)$.

Based on the same notations and reasoning as in Appendix A.2, taking $x = x_\mu^*$ in (49) and using that terms $\theta_k(f_\mu^* - d_\mu(x^k))$ and $\sum\limits_{i=1}^{k-1} \theta_i \Delta(x_\mu^*, y^i)$ are positive, we obtain:

$$\mathrm{dist}_{\mathcal{K}} \Big(G\hat{u}^k + g \Big) \leq \frac{8L_{\mathrm{d}}}{k^2} \|l^k - l^0\| \leq \frac{8L_{\mathrm{d}}}{k^2} (\|l^k - x_\mu^*\| + \|l^0 - x_\mu^*\|)$$
$$\leq \frac{8L_{\mathrm{d}}R_{\mathrm{d}}}{k^2} + 8\sqrt{\frac{3L_{\mathrm{d}}\delta}{k}}. \tag{50}$$

Further, we derive sublinear estimates for primal suboptimality. First, note that:

$$\Delta(x, y^k) = \mathcal{L}_\mu(u^k, y^k) + \langle \nabla_x \mathcal{L}_\mu(u^k, y^k), x - y^k \rangle - d_\mu(x)$$
$$= \mathcal{L}_\mu(u^k, y^k) + \langle Gu^k + g, x - y^k \rangle - d_\mu(x) = \mathcal{L}_\mu(u^k, x) - d_\mu(x).$$

Summing on the history and using the convexity of $\mathcal{L}_\mu(\cdot, x)$, we get:

$$\sum_{i=1}^{k-1} \theta_i \Delta(x, y^i) = \sum_{i=1}^{k-1} \theta_i (\mathcal{L}_\mu(u^i, x) - d_\mu(x))$$
$$\geq S_k^\theta \Big(\mathcal{L}_\mu(\hat{u}^k, x) - d_\mu(x) \Big) = \theta_k^2 \Big(\mathcal{L}_\mu(\hat{u}^k, x) - d_\mu(x) \Big). \tag{51}$$

Using (51) in (49), $\frac{\sum_{i=1}^{k-1}\theta_i^2}{S_k^\theta} \leq \max_{1\leq i\leq k-1} \theta_i \leq k-1$ and dropping term $\frac{L_{\rm d}}{2}\|l^k - x\|^2$, we have:

$$\mathcal{L}_\mu(\hat{u}^k, x) - d_\mu(x^k) \leq \frac{L_{\rm d}}{\theta_k^2}\|x^0 - x\|^2 + 3k\delta. \tag{52}$$

Choosing the multiplier $x = 0$, we observe that $\mathcal{L}_\mu(\hat{u}^k, 0) \geq f(\hat{u}^k)$ and $d_\mu(x) \leq f^* + \frac{\mu}{2}D_U^2$ for all $x \in -\mathcal{K}^*$. Then, combining this observations with (52) leads to:

$$f(\hat{u}^k) - f^* \leq f(\hat{u}^k) - d_\mu(x^k) \overset{(52)}{\leq} \frac{4L_{\rm d}\|x^0\|^2}{k^2} + \frac{\mu}{2}D_U^2 + 3k\delta \quad \leq \frac{4\|G\|^2 R_d^2}{\mu k^2} + \frac{\mu}{2}D_U^2 + 3k\delta.$$

We choose the optimal smoothing parameter by minimizing the above expression over μ and obtain: $\mu(k) = \frac{2^{3/2}\|G\|R_{\rm d}}{D_U k}$. Replacing this value in the above estimates, we obtain:

$$f(\hat{u}^k) - f^* \leq \frac{2^{3/2}\|G\|R_{\rm d}D_U}{k} + 3k\delta.$$

Also, taking $L_{\rm d} = \frac{\|G\|^2}{\mu(k)}$ in the feasibility gap (50), we get the estimate on infeasibility: $\text{dist}_{\mathcal{K}}\left(G\hat{u}^k + g\right) \leq \frac{2^{3/2}\|G\|D_U}{k} + 2\left(\frac{\|G\|D_U\delta}{R_{\rm d}}\right)^{1/2}$. On the other hand, we have:

$$\begin{aligned} f^* &= \min_{u\in U, s\in\mathcal{K}} f(u) + \langle x^*, Gu + g - s\rangle \leq f(\hat{u}^k) + \langle x^*, G\hat{u}^k + g - \left[Gu^k + g\right]_{\mathcal{K}}\rangle \\ &\leq f(\hat{u}^k) + \frac{2\|G\|D_U R_d}{k} + 2\left(\delta\|G\|D_U R_d\right)^{1/2}, \end{aligned}$$

which proves the statements of the theorem. ■

References

[1] N. Aybat and G. Iyengar, *An Augmented Lagrangian Method for Conic Convex Programming*, Working paper, \protect\vrulewidth0pthttp://arxiv.org/abs/1302.6322 , 2013.

[2] A. Beck and M. Teboulle, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM Journal Imaging Science, 2(1): 183–202, 2009.

[3] R. Bot and C. Hendrich, *A variable smoothing algorithm for solving convex optimization problems*, TOP, 23(1): 124–150, 2015.

[4] R. Bot and C. Hendrich, *On the acceleration of the double smoothing technique for unconstrained convex optimization problems*, Optimization, 64(2): 265–288, 2015.

[5] A. Beck and M. Teboulle, *Smoothing and first order methods: a unified framework*, SIAM Journal on Optimization, 22(2): 557–580, 2012.

[6] O. Devolder, F. Glineur and Yu. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming, 146: 37–75, 2014.

[7] O. Devolder, F. Glineur and Yu. Nesterov, *Double smoothing technique for large-scale linearly constrained convex optimization*, SIAM Journal on Optimization, 22(2): 702-727, 2012.

[8] G. Lan, Z. Lu and R. Monteiro, *Primal-dual first order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming*, Mathematical Programming, 126: 1–29, 2011.

[9] G. Lan and R. Monteiro, *Iteration-complexity of first order penalty methods for convex programming*, Mathematical Programming, 138: 115–139, 2013.

[10] G. Lan and R. Monteiro, *Iteration-complexity of first order augmented Lagrangian methods for convex programming*, Mathematical Programming, DOI 10.1007/s10107-015-0861-x, 2015.

[11] R. Monteiro and B. Svaiter, *On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean*, SIAM J. Optimization, 20(6): 2755–2787, 2010.

[12] I. Necoara and V. Nedelcu, *Rate analysis of inexact dual first order methods: application to dual decomposition*, IEEE Transactions on Automatic Control, 59(5): 1232–1243, 2014.

[13] I. Necoara and A. Patrascu, *Iteration complexity analysis of dual first order methods for conic convex programming*, Optimization Methods and Software, 31(3): 645–678, 2016.

[14] I. Necoara, A. Patrascu, F. Glineur, *Complexity certifications of first order inexact Lagrangian and penalty methods for conic convex programming*, Technical Report, University Politehnica of Bucharest, 2015, https://arxiv.org/abs/1506.05320

[15] I. Necoara and J.A.K. Suykens, *Application of a smoothing technique to decomposition in convex optimization*, IEEE Transactions on Automatic Control, 53(11): 2674–2679, 2008.

[16] V. Nedelcu, I. Necoara and Q. Tran-Dinh, *Computational Complexity of Inexact Gradient Augmented Lagrangian Methods: Application to Constrained MPC*, SIAM Journal on Control and Optimization, 52(5): 3109–3134, 2014.

[17] A. Nedic and A. Ozdaglar, *Approximate Primal Solutions and Rate Analysis for Dual Subgradient Methods*, SIAM Journal on Optimization 19(4): 1757–1780, 2009.

[18] A. Nemirovski, *Prox-method with rate of convergence $\mathcal{O}(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems*, SIAM Journal on Optimization, 15(1): 229–251, 2004.

[19] Yu. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming, 103: 127–152, 2005.

[20] Yu. Nesterov, *Dual extrapolation and its applications to solving variational inequalities and related problems*, Mathematical Programming, 109: 319–344.

[21] Yu. Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming, 140: 125–161, 2013.

[22] Yu. Nesterov, *Subgradient methods for huge-scale optimization problems*, Mathematical Programming, 146: 275–297, 2014, www.imtlucca.it/embopt-14/Slides/nesterov.pdf.

[23] Q. Tran-Dinh, I. Necoara and M. Diehl, *Fast inexact distributed optimization algorithms for separable convex optimization*, Optimization, 65(2): 325356, 2016.

[24] Q. Tran-Dinh, V. Cevher, *A primal-dual algorithmic framework for constrained convex minimization*, Technical report, 2014, \protect\vrulewidth0pthttp://arxiv.org/abs/1406.5403 .

[25] R.T. Rockafellar and R. Wets. *Variational Analysis*, Springer, 1998.

[26] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*, SIAM Journal on Optimization (submitted), 2008.

[27] A. Yurtsever, Q. Tran-Dinh and V. Cevher, *Universal Primal-Dual Proximal-Gradient Methods*, Technical report, 2015, \protect\vrulewidth0pthttp://arxiv.org/abs/1502.03123 .